DOCUMENT RESUME

ED 314 426 TM 014 144

AUTHOR	Rudner, Lawrence M., Ed.; Conoley, Jane Close; Plake, Barbara S.
TITLE	Understanding Achievement Tests: A Guide for School Administrators.
INSTITUTION	American Institutes for Research, Washington, DC.; Buros Inst. of Mental Measurement, Lincoln, NE.; ERIC Clearinghouse on Tests, Measurement, and Evaluation, Washington, DC.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	ISBN-0-89785-215-X
PUB DATE	Oct 89
CONTRACT	RI88062003
NOTE	169p.; For three ERIC Digests extracted from this document, see TM 014 145-147.
PUB TYPE	Guides - Non-Classroom Use (055) -- Information Analyses - ERIC Information Analysis Products (071)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	*Achievement Tests; Administrator Role; *Administrators; Elementary Secondary Education; School Districts; School Personnel; Scores; *Standardized Tests; Testing Problems; Test Interpretation; Test Results; *Test Use

ABSTRACT

Current information about tests and testing procedures is provided for school district staff, particularly in districts without specially trained testing directors. Practical information is given about selecting and administering tests and about reporting results effectively. This guide opens with a discussion of the basic principles of testing. The various types of district-level tests are described, and different types of test scores are presented. The advantages and limitations of certain types of tests and scores are reviewed. The viewpoints of measurement experts on important issues in testing are expressed in the following chapters: (1) "Common Misuses of Standardized Tests" (Eric Gardner); (2) "Preparing Students To Take Standardized Achievement Tests" (William A. Mehrens); (3) "Matching Your Curriculum and Standardized Tests" (Jane C. Conoley); (4) "Using Customized Standardized Tests" (Paul L. Williams); (5) "Interpreting Test Scores for Compensatory Education Students" (Gary Echternacht); and (6) "Working with the Press" (Allan Hartman). Four additional discussions are appended: "Finding Information about Standardized Tests' (Lawrence M. Rudner and Kathryn Dorko); Organizations That Provide Test Information" (Ronald T. C. Boyd); "Putting Test Scores in Perspective: Communicating a Complete Report Card for Your Schools" (M. Kevin Matter); and "Major Achievement Tests and Their Characteristics" (Northwest Regional Education Laboratory). Names and addresses of major test publishers, and a glossary of testing terms are also included. (SLD)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

LAWRENCE M. RUDNER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Understanding Achievement Tests:

A Guide for School Administrators

Compiled by

ERIC Clearinghouse on Tests, Measurement, and Evaluation
American Institutes for Research

Buros Institute of Mental Measurements
University of Nebraska-Lincoln

American Institutes for Research
3333 K Street, NW / Washington, DC 20007

BEST COPY AVAILABLE

Understanding Achievement Tests:

A Guide for School Administrators

Edited by

Lawrence M. Rudner / ERIC Clearinghouse for Tests, Measurement, and Evaluation

Jane Close Conoley / Buros Institute of Mental Measurements

Barbara S. Plake / Buros Institute of Mental Measurements

Published by:
The ERIC Clearinghouse on Tests, Measurement, and Evaluation
American Institutes for Research
3333 K Street, NW
Washington, DC 20007

Printed in the United States of America

This publication was prepared, in part, with funding from the Office of Educational Research and Improvement (OERI), U.S. Department of Education, under contract R-88-062003. The opinions expressed in this report do not necessarily reflect the positions or policy of OERI or the Department of Education.

Library of Congress Cataloging-in-Publication Data

Understanding Achievement Tests: A guide for school administrators / edited by Lawrence M. Rudner, Jane Close Conoley, Barbara S. Plake.
168 p.
"October 1989"
"Compiled by ERIC Clearinghouse on Tests, Measurement, and Evaluation, American Institutes for Research; Buros Institute of Mental Measurements, University of Nebraska-Lincoln" -- Cover.
Includes bibliographical references.
ISBN 0-89785-215-X
1. Achievement tests --United States--Interpretation. 2. Educational tests and measurement. I. Rudner, Lawrence M. II. Conoley, Jane Close. III. Plake, Barbara S. IV. ERIC Clearinghouse on Tests, Measurement, and Evaluation. V. Buros Institute of Mental Measurements.

LB3060.3.U53 1989
371.2'64'0973--dc20 89-27736
CIP

ISBN 0-89785-215-X
First Printing October 1989

Preface

Every year, millions of standardized achievement tests are administered to students across the country. But too often, tests are selected, administered, and reported on without serious scrutiny of the testing program and its results. The purpose of *Understanding Achievement Tests: A Guide for School Administrators* is to give school district staff -- particularly in districts without specially trained testing directors -- current information about tests and testing procedures. Our hope is that this information will lead to more careful test selection, adherence to correct administration procedures, accurate scoring and reporting, and the appropriate use of test results. When test results are accurate and credible, assessment can become a powerful instrument for improving school and student performance.

Understanding Achievement Tests was initiated partly as a result of a recent state-by-state survey of achievement test results. Conducted by John Jacob Cannell, a physician from West Virginia, the survey found that all states and about 90 percent of the sampled school districts, were surpassing the national norms for performance -- an apparent contradiction to the definition of a national norm. We now understand some of the reasons for this seemingly inflated level of test performance nationwide.

While evidence shows improved academic achievement in some cases, for instance among young minority-group children, many experts attribute the inflated scores observed by Cannell to tests being administered improperly, "teaching to the test," and outdated instruments being used.

But regardless of the cause, the unrealistically impressive scores brought to light the need for better use and greater understanding of tests. In response to that need, this guide is designed to help school administrators identify and correct poor testing procedures and inappropriate uses of test results.

Understanding Achievement Tests provides practical information about selecting and administering tests, as well as about reporting results effectively. Two companion volumes -- *A Guide to the Use of Reading Tests* and *A Guide to the Use of Mathematics Tests* -- provide detailed descriptions of the content and features of the nation's most widely used achievement tests.

Understanding Achievement Tests is a joint project of the ERIC Clearinghouse on Tests, Measurement, and Evaluation and the Buros Institute of Mental Measurements. Both institutions strive to provide accurate, objective, and current information about tests and to improve the overall practice of measurement. We hope that *Understanding Achievement Tests* contributes significantly to those goals.

Robert M. Stonehill, Director
Educational Resources Information Center (ERIC)
U.S. Department of Education

Foreword

Educational accountability is a major theme for President Bush and Secretary Cavazos, and it ranks among the top education issues in state legislatures. In fact, in this year's survey, state education-committee chairpersons ranked accountability second only to school finance. Several years ago, in *Time for Results,* the National Governors' Association touted a particular brand of accountability: the governors said they were willing to regulate schools less in exchange for better results in student learning.

By July 1987, accountability had already gathered considerable steam. The Office of Educational Research and Improvement (OERI) formed a State Accountability Study Group to examine developments in the vanguard of the accountability movement in 10 states. In the course of producing its report, *Creating Responsible and Responsive Accountability Systems* (September 1988), the Study Group commissioned the Council of Chief State School Officers (CCSSO) to do a 50-state accountability survey. Among the results, which are reported in OERI's *Measuring Up: Questions and Answers About State Roles in Educational Accountability* (November 1988), CCSSO found that 25 states have policies whereby various measures of student performance -- measures that generally include standardized test scores -- trigger consequences (rewards, sanctions, even takeovers) for schools or school districts.

Since that report, accountability has continued to gain momentum, and the quest for it has led to innovative policies in a number of states. At least four states -- California, Illinois, South Carolina, and West Virginia -- now issue "report cards" on the performance of every school in the state.

Recently, North Carolina became the first state to waive regulations for high-performing schools and Ohio is considering a similar measure. As of this writing, a gubernatorial commission in Maryland is weighing a policy that would require every school in the state to become "accredited" through periodic inspections of various educational indicators, including test scores.

This spiraling emphasis on test scores is a visible effect of the seismic shift in the debate about education in the 1980s. In the past, people talked about inputs and allocations -- equality of resources. Today when people talk about schools, the focus is on outcomes; they want to know how much and how well students are learning.

This year Rand Corporation reported on six urban school systems. The report underscores the importance of this philosophical shift. All six school systems face problems, yet according to the report, each had shown signs of improvement. Notably, all six superintendents had "promised concerted action to raise test scores" and had "published test score results for all schools and ethnic groups."

How did that help improve the schools? While there is no simple answer, one report makes this point perfectly clear: insistence on "regular and complete publication of student test scores by school and by race . . . pays off in terms of community support."

Community support is vital to improving any school, which is one reason that standardized testing ought to become a standard tool in every school leader's toolkit. However, if school leaders are going to use this tool effectively, they must view it as a process -- a process that involves three fundamental, interrelated tasks.

First, school leaders must **be clear about the purposes** of measuring their students' achievement. What kinds of decisions will be made on the basis of test results? To make those decisions, what kinds of data are necessary? At what levels must those data be disaggregated -- at the

individual student level, for particular groups of students, or for individual classrooms? How closely do they want to align the tests to curricular objectives?

Such questions help piece together a picture of the data that are needed -- a picture that is instrumental to the second task, **shopping for a test**. We are happy to report that some commercial testing companies may tailor or "customize" tests to a district's particular needs, thus increasing the alignment between the test and curricular objectives.

After tests have been administered and scored comes the third task -- **determining what the test results mean**. How do this year's scores compare to those of previous years? Why did certain groups of students score lower than others? What instructional strengths and weaknesses are revealed in the data? What policy changes appear to be needed? These are some of the questions that school leaders must answer and explain for a variety of audiences including parents, staff, School Boards, and others. Those answers must translate into a clear course of action for improving school performance and boosting student learning. That, finally, is the aim of this guide. It is also one reason why we see standardized testing as central to school leadership -- indeed, as one of the ultimate tests of school leadership. We hope that this guide helps more school leaders, as well as School Board members, teachers, parents, and others, to understand the uses and limitations of standardized testing and subsequently to harness the process as a force for strengthening teaching and learning, not only in schools but in homes and communities across the country.

Bruno V. Manno
Acting Assistant Secretary
Office of Educational Research and Improvement
U.S. Department of Education

Kirk Winters
OERI Associate
Office of Educational Research and Improvement
U.S. Department of Education

Contents

Contents

About this guide 1
What's the purpose of this guide? 1
What's in this guide? 2

A testing primer

Basic testing principles 7
What types of achievement tests are there? 9
What types of test scores are there? 19
How should you use test scores? 39

Experts' views on testing

Common misuses of standardized tests *by Eric Gardner* 47

Preparing students to take standardized achievement tests *by William A. Mehrens* 53

Matching your curriculum and standardized tests *by Jane Close Conoley* 61

Using custom-made standardized tests *by Paul L. Williams* ... 69

Interpreting test scores for compensatory education students *by Gary Echternacht* 77

Working with the press *by Allan Hartman* 85

Appendices

Finding information about standardized tests *by Lawrence M. Rudner and Kathryn Dorko* 107

Organizations that provide test information *by Ronald T.C. Boyd* 115

Putting test scores in perspective: communicating a complete report card for your schools *by M. Kevin Matter* 121

Major achievement tests and their characteristics *by the Northwest Regional Education Laboratory* 131

Names and addresses of major test publishers 151

A glossary of testing terms 153

Index

About this guide

What's the purpose of this guide?

Many people think that standardized tests are impartial indicators of how well the educational process works, how it may be improved, and how students are progressing. Administrators, parents, taxpayers, and members of School Boards expect testing programs to give accurate information that they can use to evaluate the health of their school systems. Whether a testing program can live up to these expectations depends on how well it has been designed and implemented and how well the results have been reported to various audiences.

Understanding Achievement Tests helps you, as a school administrator, deal with the tasks of selecting and using test results to accurately gauge and report educational achievement and progress. Although a booklet about standardized testing could cover many relevant topics, the focus of this guide is quite specific.

This guide emphasizes interpreting and reporting the results of **standardized norm-referenced tests** -- usually the most visible and, ironically, the least understood type of achievement test.

Understanding Achievement Tests is designed to help you

- understand basic testing principles so that you can determine what student achievement data your district needs,

- examine your testing practices so that you can match your purpose for assessing students with the appropriate testing instruments, and

- work with the press so that you can communicate the meaning and implications of your test results effectively and accurately.

This guide does not replace a good course in testing principles. In introducing you to some important concepts, its emphasis is on practical rather than theoretical aspects of standardized norm-referenced tests.

Throughout this guide, we give you charts and summaries to stimulate further thought and discussion. We encourage you to copy pages of this guide and to discuss its contents with other people in your school district.

What's in this guide?

We start with a non-technical primer, especially designed for busy school administrators. In this section of the guide, we explain the various types of district-level tests and the different types of test scores. We outline the advantages and limitations of certain types of tests

and scores, give you concrete examples and summaries, and make recommendations about how you can use different types of tests and test scores.

In the second section of the guide, we show you the viewpoints of several measurement experts who discuss important issues that confront you as you make decisions about your testing program.

Here's an overview of the issues that our experts discuss:

Common misuses of standardized tests

Eric Gardner of Syracuse University explains how standardized tests can be misused inadvertently. He tells of the need to examine testing manuals and test items to ensure that the test you choose suits your purpose. He also discusses measurement error and how that can affect your interpretation of test scores.

Preparing students to take standardized achievement tests

William Mehrens of Michigan State University points out that instruction to prepare students for standardized tests can vary from general instruction on district objectives to outright teaching to the test.

Few people will deny that students must be test-wise. For instance, students need to know how to fill out multiple-choice questions on standardized tests and when

to guess. Test preparation instruction is much more controversial, however, as it approaches teaching directly to the test. Mehrens discusses how teaching to the test affects test scores and what you can infer from test scores.

Matching your curriculum and standardized tests

Jane Close Conoley of the Buros Institute of Mental Measurements talks about selecting tests for the information they provide. When publishers develop tests, they examine text guides and district curriculum guides to determine which skills to test in each grade level. As a result, the content of a given test reflects that particular publisher's judgments about common curricula. Therefore, different publishers develop somewhat different tests.

In terms of national norms, these differences are minor because each test is normed on its own national sample and most nationally normed tests give fairly even coverage to important curriculum components. In terms of local test scores, however, these differences can have important consequences.

Conoley explains that before you select a test, you must decide if you want to evaluate your school's program or if you want to know how your school's students compare to a national sample of students. Although these purposes do not always conflict, you will meet your goals best by carefully planning and selecting a test before you administer it. Conoley details the steps you should take when you select a test for your district.

Using custom-made standardized tests

Paul L. Williams of CTB McGraw-Hill, discusses an important change in standardized tests -- using custom-made standardized tests to reduce testing time, to increase the relevance of your curriculum, and to develop greater confidence in the national comparative information.

Williams introduces you to several model tests that are used throughout the country and explains the advantages and disadvantages of each. He also explains that you should be concerned about the norm-validity of your district's test scores and how norm-valid scores can help you reflect changes in achievement.

Interpreting test scores for compensatory education students

Gary Echternacht of the Educational Testing Service offers you advice on avoiding four inappropriate practices that administrators sometimes follow when they select students and interpret test scores for compensatory education programs.

Echternacht explains how using test scores to select students, giving out-of-level tests, misinterpreting the term *grade level*, and failing to differentiate degree of error in individual and group scores can work against you as you try to develop a sound compensatory education program.

Working with the press

In a comprehensive, how-to chapter, Allan Hartman of the Massachusetts State Department of Education gives specific, practical advice about working with reporters. He gives you checklists of information that you need to compile, tips for building effective relationships with the press, and an annotated sample press release that you can use as a model.

A testing primer

Basic testing principles

Lawrence M. Rudner, ERIC/TM

What types of achievement tests are there?

Achievement tests can be put into two broad categories:

- **Norm-referenced tests** describe a student's performance in relation to the performance of a group of students.

- **Criterion-referenced tests** describe a student's mastery of particular skills.

By themselves, these labels are unimportant. Many tests and testing programs properly incorporate aspects of both types of tests. For instance, many criterion-referenced tests have been normed and many norm-referenced tests permit content-based interpretation.

Further, it is not easy to tell one type of test from the other simply by looking at the test's items. However, norm-referenced tests are constructed differently from

criterion-referenced tests and, consequently, their primary strengths and limitations are different.

Before you can select the particular type of test that is appropriate for your school, you must understand your purposes for testing. In the following pages, we explain the differences between norm-referenced and criterion-referenced tests to help you make informed choices.

Norm-referenced tests

Norm-referenced tests help you compare **one** student's performance with the performances of a **large group** of students. Norm-referenced tests are designed to make fine distinctions between students' performances and accurately pinpoint where a student stands in relation to a large group of students.

Norm-referenced tests are usually developed by commercial test companies and, typically, many schools use the same test. Among the better known norm-referenced tests are the Iowa Test of Basic Skills, the Stanford Achievement Test, and the California Achievement Test.

How are norm-referenced tests created?

When developers create norm-referenced tests, they carefully survey existing curricula so that they can write test items to reflect the material that is taught in most schools. Based on this analysis, they prepare detailed test specifications, or test blueprints, that outline the curricular

objectives that will be measured and the number of items that will be used to assess each objective. These objectives then guide the developers in writing the test items.

To ensure that the final test has a sufficient number of high-quality items in each curricular area, developers usually pilot test items on a sample of students using two to three times as many items as is planned for the final version of the test. In developing tests that are going to be used nationally, these "tryout samples" closely match the U.S. student population in terms of such variables as community size, geographic region, family income, years of parental schooling, and nationality.

Based on the results of the pilot test, developers retain only the test items that meet certain statistical standards. A good item

- **generates consistent responses,**
- **is not biased against any ethnic or gender group,** and
- **measures the desired learning objectives**

To use the Primary II battery of the Stanford Achievement Test as an example, 2,565 items were piloted and 1,326 items were retained for the three final forms.

One of the most important criteria for deciding whether to retain a test item is how well that item contributes to the variability of test scores. Good test developers compose items that encourage variability. They create items that are neither too easy nor too hard and then use the item tryout to confirm their decisions.

An effective norm-referenced test is able to make fine distinctions among students' abilities. It can accurately rank students from highest to lowest ability. While they may be closely related to learning outcomes, items that are too easy or too difficult do not contribute to the variability of test scores and usually will be eliminated.

After developers select the items for a test, they develop test norms and normative test scores, such as grade equivalent scores and percentiles. These norms provide a means to compare the performance of one student or group of students with the performance of a specified reference group. While it is possible to have several reference groups, most standardized achievement test batteries use a representative sample of the U.S. population of school children as the benchmark. Most publishers will also compute district-level norms.

In context, these norms have meaning for most school systems. The norms describe the typical performance of U.S. students on these items at the time the norms were developed.

Several publishers now create *custom-developed norm-referenced tests*. These customized tests are based on your local curricular objectives and come with national norms. As we explain in greater detail in the next chapter, these norms are valid only under certain circumstances.

What are the advantages of norm-referenced tests?

- They allow you to analyze the general progress of large groups of students.
- They give you a basis for examining an individual student's general performance.

What are the limitations of norm-referenced tests?

- They are inappropriate for following an individual student's progress in specific skills.
- They are insufficient for diagnosing a student's specific strengths or weaknesses within a given subject area.
- They may be inappropriate for your district if specific features of your curriculum or of your students are not represented in the test.
- They assess a relatively narrow range of desired educational outcomes.

- They provide a limited number of items to measure each objective.

- The norms quickly become outdated.

Criterion-referenced tests

Criterion-referenced tests help you determine which specific skills individual students have mastered. Detailed information about a student's skills can help you make decisions about that student and about your programs. This kind of information can help you focus your instruction to concentrate on specific weaknesses of your students. Information about groups of students can also be useful in evaluating whether a program was successful in helping students achieve specific objectives.

Criterion-referenced tests are usually developed to reflect the skills taught in a local school district. Because curricula vary, different districts typically use different tests.

Many larger school districts develop their own criterion-referenced tests. In addition, many districts with criterion-referenced tests make their items available to other local districts to help them develop their own tests. Also, most commercial test publishers can help you develop a custom-made criterion-referenced test.

Because criterion-referenced tests are designed to reflect the local curriculum, you must invest a great deal of time in defining the objectives to be tested. You must then

write many items to reflect these specific objectives. Typically, criterion-referenced tests cover more skills and have more items per skill than norm-referenced tests; thus, they are often much longer than norm-referenced tests.

Variability of test scores is not as important in criterion-referenced tests as it is in norm-referenced tests. With criterion-referenced tests, the goal is not to rank students, but to have scores that reflect whether students have mastered certain skills. If a skill is targeted in the school, then the criterion-referenced test should contain items to measure mastery of that skill.

Most norm-referenced tests can also provide some criterion-referenced information. By clustering items that measure a common objective, test publishers can report whether particular skills have been mastered. These data can be extremely valuable when planning instruction or evaluating programs. However, criterion-referenced tests which are designed specifically for this purpose give you more detailed and more accurate information than norm-referenced tests that are converted to give criterion-referenced data.

What are the advantages of criterion-referenced tests?

- They measure whether your district has attained its curricular objectives.
- They are often developed from programs or courses that are taught in local schools.

- They may be appropriate for diagnosing your students' strengths and weaknesses within a given subject area.

- They help you plan instructional programs.

What are the limitations of criterion-referenced tests?

- They do not usually provide meaningful norms.

- They can be expensive to develop.

- You must revise them periodically to reflect your current objectives.

- They require a great deal of testing time.

Table 1 enumerates the appropriate uses of both norm-referenced and criterion-referenced tests.

Table 1. Appropriate Uses of Norm-referenced and Criterion-referenced Tests [1]

Purpose	Test	Examples	Primary users
To compare achievement of local students to achievement of students in the nation, state, or other districts in a given year.	NRT	A comparison of achievement of local schools' 3rd graders to achievement of 3rd graders throughout the nation.	Central office, (including school boards), parents
To compare achievement of subgroups of local students to achievement of similar subgroups in the nation, state, or other districts in a given year.	NRT	A comparison of achievement of local black to the achievement of black students throughout the nation.	Central office
To compare achievement of one local school's student subgroup (e.g. sex, race, or age) to achievement of another such subgroup in a given year to determine the equity of educational outcomes.	NRT	A comparison of achievement of black and white students in local schools to determine and monitor any gap in achievement.	Central office, principals
To assess the extent to which students in a single grade level (at district, building, or classroom level) have mastered the essential objectives of the school system's curriculum.	CRT	A comparison of difference between results of September and May criterion-referenced tests to determine the extent to which 3rd graders at a given school attained 3rd grade objectives in reading.	Teachers, principals, central office
To assess the extent to which a given student is learning the essential objectives of the school system's curriculum and, subsequently, to adjust instruction for that student.	CRT	The use of the results from the September and January criterion-referenced tests as one indicator to help determine if a student is properly placed in an instructional group.	Teachers, principals, parents

[1] This chart was originally prepared by Prince George's County Maryland Public Schools.

Table 1. Appropriate Uses of Norm-referenced and Criterion-referenced Tests (continued)

Purpose	Test	Example	Primary Users
To track achievement of cohort of students through the system or area to determine the extent to which their achievement improves over time.	CRT	An examination of progress of all 3rd graders in system, administrative area, or school from one year to the next.	Central office, principals
To track achievement of cohort of students in a given school to determine the extent to which they learn essential objectives of school system's curriculum as they go from grade to grade.	CRT	The use of May criterion-referenced tests (or perhaps gains from September to May), to follow the progress of children over time in terms of the extent to which they learned the curriculum from one year to another.	Principals, teachers

What types of test scores are there?

Different types of scores provide different types of information and serve different purposes. You must understand the different types of scores before you can select scores that are most appropriate for your needs.

In this section, we define these types of test scores:

- *raw scores,*
- *total percentage correct scores,*
- *object mastery scores,*
- *percentile scores,*
- *stanine scores,*
- *grade equivalent scores,*
- *standard scores, and*
- *normal curve equivalent scores*

and explain the advantages and disadvantages of each. In the next section, we discuss how to use them.

Remember that test scores reflect only what was measured on a particular test. For example, scores on the Iowa Tests of Basic Skills (ITBS) test of mathematics achievement reflect only the combination of skills tested by the ITBS. Scores on other mathematics tests are not necessarily comparable.

Raw scores

Raw scores indicate the number of items a student answers correctly on a test. For students who take the same test, it makes sense to compare their raw scores. If one third grade student answers 12 of 25 items correctly and another answers 16 correctly, then the second student knows the content better than the first.

Because the number of items varies between tests and because tests vary in difficulty, raw scores have little value in making comparisons from one subject to another. Suppose a third grade student answers 12 out of 25 items correctly on a mathematics test and 16 out of 25 items on a reading test. Some people may assume that the student is better in reading than in mathematics. However, we really know nothing about relative performance in the two different areas because the mathematics test may be much harder than the reading test.

How are raw scores distributed?

As an example of how raw scores are usually distributed over the population, let's look at a national sample of 2,000 students.

If you give a 25-item mathematics test to a large number of students, you will typically find the largest number of students have scores around the average and the number of students with a given raw score decreases the further you get from the mean.

Figure 1 illustrates a hypothetical number of students with each test score.

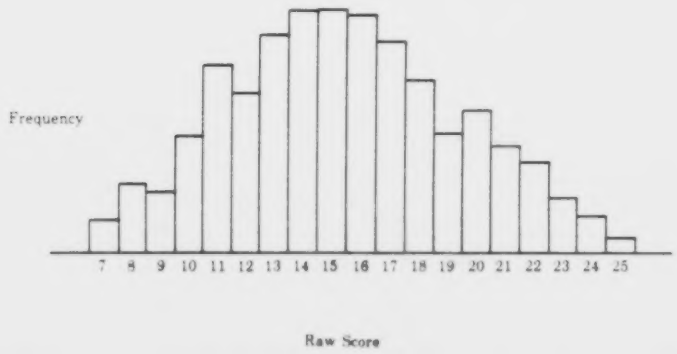

Figure 1. Raw Scores

The distribution of test scores shown in Figure 1 can be modeled mathematically using the familiar bell-shaped "normal" curve.

In the normal curve shown in Figure 2, the *y* axis shows the relative proportion of students and the *x* axis shows total raw score. The curve shows the predicted proportion of students who would have a given total score.

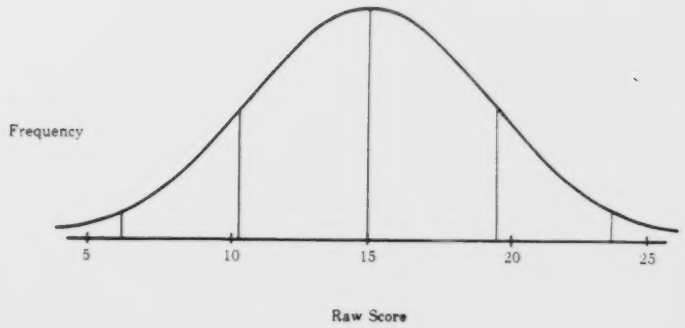

Figure 2. The Normal Curve

The normal curve is only a mathematical model that shows a relationship between two variables -- **test scores** and **proportion of students**. Actual scores never perfectly match the model. Nevertheless, the model is close to reality and gives good practical results. The same relationship between test scores and proportion of students holds for a wide number of tests. Test developers use the model of the normal curve in developing and norming tests. In this guide, we use it to show similarities between different types of normative test scores -- test scores that describe individual student performance in comparison to the actual performance of a large group of students.

Two statistics are helpful in discussing test score distributions:

- the *mean* and
- the *standard deviation*.

The ***mean*** is frequently called the *average* score. You compute the mean by adding all the scores then dividing the sum by the total number of scores.

A *deviation* score is *how far away the score is from the mean*. For example, on a test with a mean of 15, a score of 20 deviates 5 points from the mean. The deviation score alone does not tell you whether this is a big difference or not. Rather, the *standard deviation* gives you a framework for interpreting this test score variability. You compute the standard deviation by taking the square root of the averaged, squared deviation. You can interpret standard deviation as the average distance that the scores deviate from the mean.

What are the advantages of raw scores?

- They are easy to compute.

- One of the most accurate ways to analyze a student's gains in achievement is to compare the raw scores from two administrations of the same test.

What is the limitation of raw scores?

Raw scores do not contain a frame of reference for indicating how well a student is performing.

Total percent correct scores

Total percent correct scores tell you the percentage of items that a student answers correctly out of the total number of items on a test. Like raw scores, total percent correct scores do not reflect varying degrees of item and test difficulty. They are of limited value in making comparisons.

Note that total percent correct scores are NOT the same as percentile scores. (We discuss percentile scores later in this section.)

What are the advantages of total percent correct scores?

- They are easy to compute.
- They adjust for differing numbers of items.

What are the limitations of total percent correct scores?

- They do not adjust for differing test difficulties.

- They do not contain a frame of reference for indicating how well a student is performing.

Objective percent correct scores

Objective percent correct scores tell you the percent of the items measuring a single objective that a student answers correctly. Because objectives and items can vary in difficulty, this score is of limited value for determining whether a student has mastered a learning objective.

You should interpret the objective percent correct score in relation to an **expected** objective percent correct. Expectations are sometimes based on curricular goals, last year's performance, or national averages.

Expectations can be used to convert objective percent correct scores to *objective mastery scores*. When the expectation is met or exceeded, the **objective is mastered**. Conversely, when the score is lower than expected, the objective is not mastered.

For example, suppose a test contains eight whole-number addition problems and a student answers seven of them correctly. That student's objective percent correct score is 87.5%. If you feel that answering six questions correctly reflects mastery, then this test score indicates that the student has mastered the objective.

What are the advantages of objective mastery scores?

- They are easy to compute.
- They adjust for differing numbers of items per objective.
- They help you diagnose specific individual strengths and weaknesses.
- They provide a skill-based approach to classroom grouping and school-based curricular emphasis.

What are the limitations of objective mastery scores?

- They require a fairly large number of items (usually more than ten) for each objective. The fewer items there are per objective, the greater is the likelihood of mistaking masters from non-masters and vice versa.
- Expectations are not always easy to define. The national average is not always a good basis for determining expectation.
- They do not indicate the degree or level of skill that the student has attained; they only indicate the status of mastery or non-mastery.

Percentile scores (ranks)

Percentile scores tell you the percent of students in the norming sample whose scores were at or lower than a given score. Percentile scores are among the most commonly reported scores and are best used to describe a student's standing in relation to the norming group at the time of testing. For example, if a student's score is in the 80th percentile, then that student scored equal to or higher than 80% of the students who took the test when the test was normed.

Note that although percentile scores are reported in increments of one hundredths, they are not completely accurate. Percentile scores are usually accurate only to the nearest six one-hundredths (.06) because of measurement error. Therefore, when you use percentiles, you should pay attention to the *confidence bands* that the test publisher provides.

Confidence bands represent the **range of scores** in which a student's true score is likely to fall. For example, although a student's score on a particular test may be at the 86th percentile, it is likely that if the student took the same test on a different day, the new score would vary slightly. Accounting for random variations, that student's true achievement may fall somewhere within a range of scores, for example, between the 83rd and 89th percentiles.

Percentile units are used to report an individual student's score; they should not be averaged to describe groups. Percentile units cannot be subtracted to compute gains because differences in percentile scores are not

constant across the entire scale. For example, getting an additional two items correct can greatly increase a percentile rank for an average student. Yet the score increase from the same two items may not result in any percentile change for students of very above average achievement. Score gains increase percentile ranks more in the middle of the range than toward the extremes. (See Figure 3.)

How are percentile scores distributed?

Figure 3 shows how percentile scores are distributed when raw scores are distributed normally. The *y* axis shows the proportion of students and the *x* axis shows the percentile score. Vertical lines have been drawn to indicate each standard deviation unit.

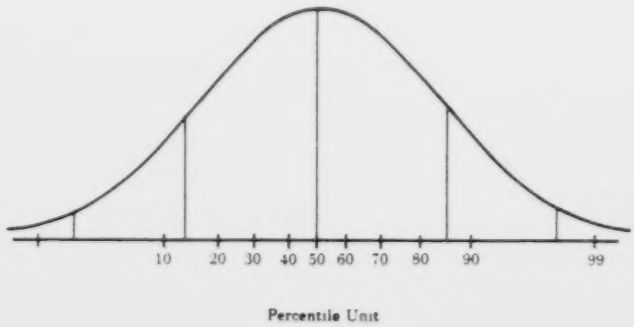

Figure 3. Percentile Score Distribution

Notice that percentiles are more "spread out" at the ends of the figure. For example, the raw score difference between the 95th and 90th percentile is greater than the difference between the 55 and 50th percentile. This happens because a student needs to answer more items correctly to move from the 90th to the 95th percentile than is necessary to move from the 50th to 55th percentile. Therefore, scores are clustered around the mean. It is because of this difference that you should not add, subtract, or average percentiles.

What are the advantages of percentile scores?

- They show how students rank in relation to the national or local average.
- They are easy to explain.

What are the limitations of percentile scores?

- They can be confused with total percent correct scores.
- They are not as accurate as they appear to be.
- They are often used inappropriately to compute group statistics or to determine gains.
- They are frequently misunderstood.

Stanine scores

Stanine is short for *standard nine*. Stanine scores range from a low of 1 to a high of 9 with:

- 1, 2, or 3 representing **below average**
- 4, 5, or 6 representing **average**
- 7, 8, or 9 representing **above average.**

If a student achieves a stanine score that is below average in a particular area, the test has revealed an area in which the student may need to improve -- or at least it reveals an area in which the student is weak when compared to other students who took the test. If the student achieves an average stanine score, the test has revealed that the student performed at the same level as most of the other students who took the test. Similarly, if the student achieves a stanine score that is above average, the test revealed that the student performed better in that area than most of the other students who took the test.

Stanines are frequently used as a basis for grouping students. For example, an advanced mathematics class may enroll students in the 9th, 8th, and sometimes 7th stanine.

How are stanine scores distributed?

Figure 4 shows how stanine scores are distributed when raw scores are distributed normally. The *y* axis shows the proportion of students and the *x* axis shows the stanine score. Vertical lines have been drawn to indicate each standard deviation unit. Stanine 5 represents 1/2 a standard

deviation (sd) around the mean. Stanines 2, 3, 4 and 6, 7, and 8 also represent the same raw score difference (1/2 sd). Stanines 1 and 9 represent all the scores below -1.75 sd and above +1.75 sd, respectively.

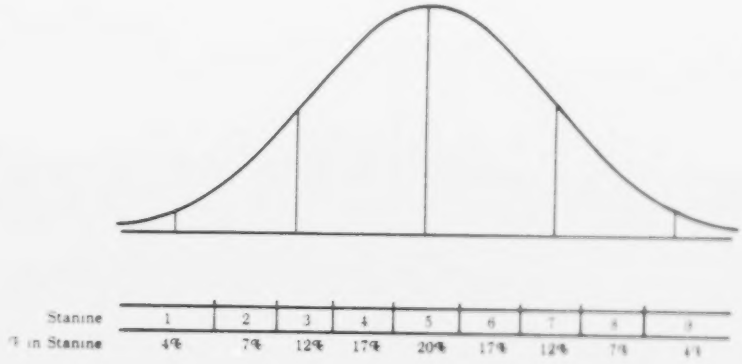

Figure 4. Stanines

District test results can be reported by showing the percent of district students who fall in each stanine compared to the national average.

What are the advantages of stanine scores?

- They show the standing of students in relation to the national or local average.
- They are easy to explain.

- They can be used to group students into ability groups.

What are the limitations of stanine scores?

- They should not be used in computing group statistics or in determining gains.
- They give only very general indications of a student's relative standing in a particular content area.

Grade equivalent scores

Grade equivalent scores use a scale based on grade levels and months to estimate how well students perform. These scores reflect the median score of students across several grade levels during the month the test was normed. For instance, the median test score for first graders in the seventh month of the school year would convert to a score of 1.7, for second graders the score would be 2.7, for third graders the score would be 3.7, and so forth.

Grade equivalent scores are often misunderstood. For example, if a fourth grader received a grade equivalent score of 7.0 on a fourth grade reading achievement test, some people may assume that the fourth grader has mastered seventh grade material. However, the score actually means that the fourth grader reads fourth grade material as well as the typical beginning seventh grader would read the same fourth grade material.

As with percentile scores, you should use grade equivalent scores only to describe a student's standing in relation to the norming group at the time of testing. You should not average grade equivalent scores to describe groups, and you should not subtract them to compute gains.

As with differences in percentile scores, differences in grade equivalent scores do not mean the same thing across the entire scale.

How are grade equivalent scores distributed?

Figure 5 shows an example of how grade equivalent scores are distributed when raw scores are distributed normally. The *y* axis shows the proportion of students and the *x* axis shows the grade equivalents. Vertical lines have been drawn to indicate each standard deviation unit.

Note that this is just an example, because grade equivalent scores are not defined by the model but rather by the actual performance on the test by students in higher and lower grade levels.

Notice that relatively few correct responses translate to large differences in grade equivalent scores for students who achieve very high and very low scores. Because of this, grade equivalent scores do not estimate group ability well and you should not use them to evaluate gains over time.

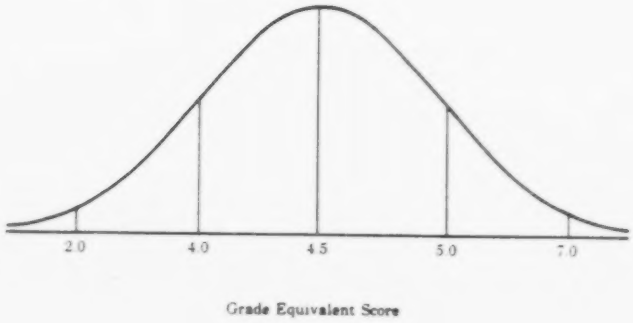

Figure 5. Grade Equivalent Scores

What is the advantage of grade equivalent scores?

Grade equivalent scores are expressed in grade-level values that are familiar to parents and teachers.

What are the limitations of grade equivalent scores?

- They are frequently misunderstood and misinterpreted.
- They have low accuracy for students who have very high or very low scores.

- They should not be used for computing group statistics or in determining gains.

Standard scores

Standard scores tell you how much students' scores deviate from a mean. Almost all of the companies that publish achievement tests will give you standard scores. However, they often use different names -- such as *growth scale values, developmental standard scores,* and *scaled scores* -- and different units to report the scores. Thus, a scaled score of 110 on one test may not be the same as a scaled scores of 110 on another.

The main advantage of standard scores is that they give you an equal interval unit of measurement. As a result, you can use them to compute summary statistics, such as averages and gains, if all the students you compare took the same test. A two-point difference between standard scores means the same difference, no matter where a students falls within the range of scores (unlike percentile and grade equivalent scores).

As we noted, the scales used for standard scores differ among test publishers and among content areas. As a result, you cannot usually use these scores to compare results on different tests.

How are standard scores distributed?

Figure 6 shows how standard scores are distributed on the a hypothetical test when raw scores are distributed normally. Here the raw scores have been translated to a scale with a mean of 100 and a standard deviation of 10. The *y* axis shows the proportion of students and the *x* axis shows the standard score. Vertical lines have been drawn to indicate each standard deviation unit.

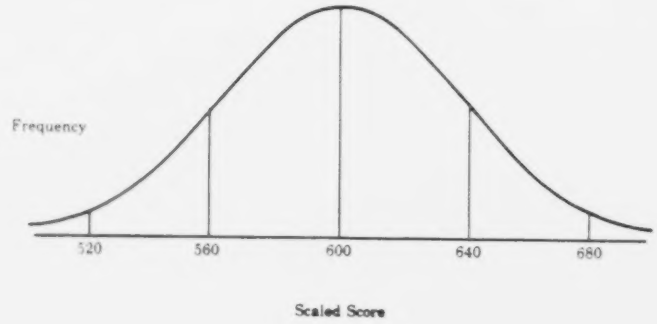

Figure 6. Standard Scores

Note that the intervals in Figure 6 are equal in size. This feature makes standard scores and scores based on standard scores the statistic of

choice when reporting group averages and changes over time.

What are the advantages of standard scores?

- They allow you to compare the achievement of students who take different levels of the same test within a test battery.
- They allow you to compare a student's achievement across subject matter.
- You can use them to compute meaningful summary statistics.
- You can use them to evaluate gains over time.

What are the limitations of standard scores?

- They do not give you information about an individual student's achievement level, unless you compare them to another value or convert them to percentile or grade equivalent scores.
- They can be confusing to parents and teachers unless they are converted to percentile scores.
- They have no intrinsic meaning, unless the scale is commonly understood because it is used frequently. For example, the Scholastic Aptitude Test for college admissions uses a

standard score with a mean of 500 and a standard deviation of 100.

Normal curve equivalent scores

Normal curve equivalent scores were originally developed to analyze and report gains in compensatory programs for educationally disadvantaged students. These scores have a mean of 50 and a standard deviation of approximately 21. This results in a scale with 99 equal interval units.

A normal curve equivalent score of 50 represents the national average of any grade level at the time of year the test was normed. A score of 30 is always the same distance below grade level, regardless of the level tested, and is twice as far below grade level as a score of 40.

Normal curve equivalent scores are similar in their range to percentile scores, but they have statistical properties that allow them to be used to compute summary statistics and gain scores.

How are normal curve equivalent scores distributed?

Figure 7 shows how normal curve equivalent scores are distributed when raw scores are distributed normally. The *y* axis shows the proportion of students and the *x* axis shows the score. Vertical lines have been drawn to indicate each standard deviation unit.

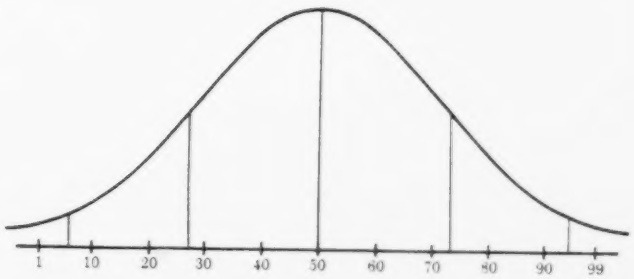

Figure 7. Normal Curve Equivalent Score

Because normal curve equivalents are a type of standard score, they have the same statistical properties as standard scores. Normal curve equivalent intervals are of equal size and these scores can be used to compute group statistics.

What are the advantages of normal curve equivalent scores?

- They allow you to compare the performance of students who take different levels or forms of the same test within a test battery.

- They allow you to draw comparisons across subject matter for the same student.

- They can be used to compute meaningful summary statistics.
- They can be used to evaluate gains over time.
- They can be used to combine data from different tests.

What is the limitation of normal curve equivalent scores?

Normal curve equivalent scores do not give you easily understood information about an individual student's achievement level, unless they are compared to another value or are converted to a percentile score.

How should you use test scores?

Interpreting norm-referenced test scores

Normative test scores -- stanines, percentiles, scaled scores, and grade equivalent scores -- measure **an individual** student's achievement in relation to the achievement of **one or more large groups** of students who took the same test. The comparison group may be composed of other students in your district or of students from a nationally representative sample. Thus, scores on norm-referenced tests are meaningful only in relationship to a comparison group.

Your school or district is not completely like the normative group. No district is. In many cases, the differences are minor and inconsequential. However, in other cases, schools can be so different that the national norms provided by the publisher do not accurately reflect school performance. Norms become less meaningful as your students and your testing program become more unlike the standardization sample.

If your students are tested at a different time of the year than the norm group was tested, the interpretation of the percentile score is unclear. For example, the CAT is normed in October. That means that you must give it in October to make your students' scores most meaningful. If you give the CAT in January, you cannot know if a student who scores in the 55th percentile is above or below average when compared to grade-level peers. (See the Appendix called *Communicating a complete report card for your school* for a list of the many ways in which your students, schools, and district may be different from the normative sample.)

Many of these differences can seriously affect your scores. This does not mean the national norms are useless; it means that you must evaluate the norms in perspective. Norms give you an index of how well students perform on certain tasks -- tasks the test publishers have identified as representing the skills taught to the comparison group at the time the test was developed.

Scores that are above average, for example, may be only above the average of students in the norm group who

were tested four years ago. They may not be above today's average.

The comparative baseline of norm-referenced tests is a powerful tool. In addition to worrying whether your Chapter 1 students are learning basic skills, for example, you probably are also interested in how well they are doing in relation to the nation. Although your students may not be like the nation at large, they are going to be competing for jobs and educational opportunities against a wide range of other students.

While national averages give you a baseline, you must establish your own expectations and goals considering your particular community and curriculum. For example, it would be somewhat misleading for you to report above average scores for a magnet school that selects students based on academic achievement. In this case, you would be better off reporting on the gains or specific achievements of the students who are in the program.

Using the right scores

If you are interested in student gains, for example, you have two psychometrically sound options. First, you can use raw scores without any transformation. However, this is only possible when the same test is given during two administrations. Second, you can use scaled scores, regardless of the examination level taken.

One of the major advantages of achievement tests is the large number of ways in which test results can be scored

and analyzed. In this section of the guide, we have described several types of test scores and outlined their advantages and limitations.

Table 2 identifies the test scores that are most appropriate for a given purpose.

Table 2. Appropriate Uses of Different Test Scores

Purpose	Use these scores	Examples	Audiences
To report what a student can and cannot do	Objective mastery scores	Identifying specific math skills mastered by each student in the class. Identifying school-wide weaknesses	Parents, teachers, principals
To report a student's performance in relation to other students	Stanine, percentile, or grade-equivalent scores	Identifying whether a student is performing as well as other students	Parents
To report a student's relative strengths and weaknesses	Stanine, percentile, or grade equivalent scores	Identifying whether a student is better in math or reading	Parents
To compute the average performance of groups of students	Scaled scores	Analyzing differences in performance between black and white students	Central office
To compute gains over time	Raw scores, scaled scores	Evaluating the effectiveness of a new reading program	Principals, central office
To aggregate data from different tests	Normal curve equivalent scores	Describing programs that cut across schools or school districts	Central office, state education agency

Experts' views on testing

In the previous sections of this guide, we introduced you to a broad array of topics about standardized testing. In the following pages, different testing experts discuss important issues that you should consider. You may want to refer to this material later when questions arise in your school district.

Common misuses of standardized tests [2]

Eric Gardner, Syracuse University

Look beyond the title of the test

Unsophisticated test users tend to accept tests' titles as accurate and complete descriptions of the variables being measured. Since titles of standardized tests must be brief, they cannot convey all the information that you must know about the kind of behavior that the test measures. All standardized tests are open to this kind of uncritical use.

Since cognitive ability has so many facets, no test can adequately measure all of them. You'll only know what is being measured if you fully understand the particular items on a given test. Furthermore, the testing situation may completely change the expected behavior. For instance, if a student who doesn't speak English or who is blind takes an aptitude test that is printed in English, that

2 Reprinted from Ability Testing: Uses, Consequences, and Controversies, 1982, with permission from the National Academy Press, Washington, DC

particular test obviously doesn't accurately measure any aspect of that student's aptitude or intelligence.

In a less obvious example, a test that is labeled "Science Achievement" may be an acceptable test to sample the science curriculum for students in a **particular** fifth grade science course, but it may fail to function as a science test for most pupils if the degree of reading difficulty is at the high school level.

A test producer's claims about an achievement or aptitude test do not mean that the test will function as such in **all circumstances** with **all pupils**. If you don't carefully examine both the test manual and the test items to determine the specific aspects of cognitive ability to be tested (such as memory, vocabulary, or type of reasoning), you can misuse the test simply because you selected an inappropriate test for your particular purpose or situation.

Understand the error of measurement in test scores

Every test score has an error of measurement. You misuse test scores or observations if you accept them as fixed, unchanging indices that contain no error.

It is impossible to say with certainty that students' observed scores give their "true" performance on the general domain about which inferences are to be made. The best that can be done is to experimentally estimate the standard error of measurement; then use that value to

set up a band within which a probability can be stated about whether the "true" score is within that band.

For example, the SAT furnishes useful data even though

- you cannot accept an SAT score of 550 as a precise measure,
- you must accept a range of scores, and
- you must then expect to be wrong a certain proportion of the time.

You can misuse test scores if you interpret a score without knowing the size of the error of measurement. In the case of most standardized test scores, the magnitude of the error is explicitly stated; it is not hidden or unknown. In fact, the errors made in grading essays have far greater -- and usually unknown -- errors of measurement.

Some people reject the notion of basing decisions on probabilistic data. However, probability estimates are involved in almost every decision we make. For example, the decision to cross a busy street at a particular instant is not made with a probability of 1.0 of doing so safely.

Don't use a single test score to make a decision

You must consider and interpret scores in the full context of the various elements that characterize students, teachers, and the general educational environment. A single test score represents only a sample from a limited domain and does not include the variety of factors that might influence that score.

For example, in making college admissions decisions, SAT scores should not be considered by themselves and are, in fact, usually weighed along with high school records and other relevant data, such as teachers' or supervisors' recommendations about motivation, leadership ability, creativity, and involvement in extracurricular activities. All of these elements can then be evaluated against socioeconomic background, social obstacles, or unusual physical demands that students must overcome to reach their current educational levels.

Understand how test scores are reported

Many people misunderstand what test scores mean. Some believe they understand raw scores or how particular raw scores are converted to total percent correct scores. However, even in this most elementary illustration, more is involved than a single number indicates. For example, 45 items answered correctly out of 50 **easy** items means something substantially different than 45 items answered correctly out of 50 **very difficult** items from the same domain.

Interpreting how raw scores are converted to percentile scores causes even more problems. The statement that "In a norm-referenced test, half the pupils must fail," doesn't tell you much about an individual student's performance. You must understand how a given score fits in with the scores of the group of students who were used to create the scale. For instance, if the group consisted of students who had high ability or unusual skills, a seemingly low percentile rank of 20 might truly indicate an excellent or even remarkable performance.

People more commonly misinterpret grade equivalent scores. A grade equivalent score is the score that 50% of the group exceeded at the specific time when the test was given. It does not represent a standard to be attained nor does it represent the grade in which the student should be placed.

Understand what tests measure

Many people confuse the information that a test score provides with the interpretation of what caused the behavior that a test score describes. A test score is a *numerical description of a sample of performance at a given point in time*. A test score gives **no information** about why students performed as reported.

Furthermore, no statistical manipulation of test data, even though combined with the best additional data, will permit more than probabilistic inferences about causation or future performance. The current reports on the decline of SAT scores are excellent examples of how difficult it is to ascribe causation to known performance. The investigating panel charged the researchers with explaining the causes of the drop in SAT scores. The researchers were able to describe the drop in scores and offer changes in test populations as a plausible partial explanation for the initial drop, but they could only speculate about the effect of other variables and the reasons for the continued drop.

Preparing students to take standardized achievement tests

William A. Mehrens, Michigan State University

As a school administrator, you know that the public often favors accountability in education and believes that holding teachers responsible for students' achievement will result in better education. Many people assume that the best data about students' levels of achievement come from standardized achievement tests. Although scores from these tests are undoubtedly useful for accountability purposes, educators recognize that such data are limited.

Teaching to the test

One major concern about standardized achievement tests is that when test scores are used to make important decisions, teachers may *teach to the test* too directly. Although teaching to the test is not a new concern, today's greater emphasis on teacher accountability can make this practice more likely to occur.

Depending on how it is done, teaching to the test can be either productive or counterproductive. Therefore, you

need to carefully consider how you prepare students to take standardized achievement tests.

At some point, legitimate teaching **to** the test can cross an ill-defined line and become inappropriate teaching **of** the test (Shepard and Kreitzer, 1987). Educators may disagree about what specific activities are inappropriate. However, it may be useful to describe a continuum and to identify several points located along it.

Seven points on the continuum

Mehrens and Kaminski (1989) suggest the following descriptive points:

1. giving general instruction on district objectives without referring to the objectives that the standardized tests measure;

2. teaching test-taking skills;

3. providing instruction on objectives where objectives may have been determined by looking at the objectives that a **variety** of standardized tests measures (The objectives taught may or may not contain objectives on teaching test-taking skills.);

4. providing instruction based on objectives (skills and subskills) that specifically match those on the standardized test to be administered;

5. providing instruction on specifically matched objectives (skills and subskills) where the practice or

instruction follows the same format as the test questions;

6. providing practice or instruction on a published parallel form of the same test; and

7. providing practice or instruction on the test itself.

Mehrens and Kaminski suggest that:

- Point 1 is always ethical and Points 6 and 7 are never ethical.

- Point 2 is typically considered ethical.

Thus, the point at which you cross over from a legitimate to an illegitimate practice on the continuum is somewhere between Points 3 and 5. The location of the point changes depending on the inferences you want to make from the test scores.

What you can infer from test scores

The only reasonable, **direct** inference you can make from a test score is the degree to which a student knows the content that the test samples. Any inference about **why** the student knows that content to that degree. . . is clearly a weaker inference. . . (Mehrens, 1984, p. 10).

Teaching to the test alters what you can interpret from test scores because it involves teaching specific content. Therefore, it also weakens the direct inference that can be reasonably drawn about students' knowledge. Rarely

would you want to limit your inference about knowledge to the specific questions asked in a specific format. Generally, you want to make inferences about a broader domain of skills.

Further complicating matters, many people wish to use test scores to draw **indirect** inferences about *why* students score the way they do. Indirect inferences can lead to weaker and possibly incorrect interpretations about school programs.

Indirect inferences cannot possibly be accurate unless the direct **inferences are made about student mastery of the content samples by the test.** Rarely does one wish to limit the inference about knowledge to specific questions in tests or even the specific objectives tested. For example, if parents want to infer how well their children will do in another school next year, they need to make inferences about the broader domain and not about the specific objectives that are tested on a particular standardized test. For that inference to be accurate, the instruction must not be limited to the narrow set of objectives of a given test. Thus, for the most typical inferences, the line demarking legitimate and illegitimate teaching of the test must be drawn between Points 3 and 4.

While in my view it is inappropriate to prepare students by focusing on the sample of objectives that happen to be tested, you can undertake appropriate activities to prepare students to take standardized tests.

Appropriate activities to prepare students

Ligon and Jones suggest that an appropriate activity for preparing students for standardized testing is:

> *one which contributes to students' performing on the test near their true achievement levels, and one which contributes more to their scores than would an equal amount of regular classroom instruction (1982, p. 1).*

Matter suggests that:

> *Ideally, test preparation activities should not be additional activities imposed upon teachers. Rather, they should be incorporated into the regular, ongoing instructional activities whenever possible (1986, p. 10).*

If you follow the suggestion by Ligon and Jones, you might spend some time teaching students general test-taking skills. These skills would help students answer questions correctly if they have mastered the objectives. Without some level of test-taking skills, even knowledgeable students could miss an item (or a set of items) because they did not understand the mechanics of taking a test.

Summary

Although the temptation exists to teach too closely to the test, teachers should not be pressured to do so. In fact, you should try to ensure that they do not do so.

The inferences you typically wish to draw from test scores are general in nature and will be inaccurate if you limit instruction to the actual objectives sampled in the test -- or, worse yet, to the actual questions on the test. However, it is appropriate to spend some instructional time teaching test-taking skills. Such skills are relatively easy to teach and should take up very little instructional time.

References

Ligon, G. D. and Jones, P. (April l, 1982). *Preparing Students for Standardized Testing: One District's Perspective*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Matter, M. K. (1986). "Legitimate Ways to Prepare Students for Testing: Being Up Front to Protect Your Behind." In J. Hall and F. Wolmut (eds.). *National Association of Test Directors 1986 Symposia*. (pp. 10-11). Oklahoma City, OK: Oklahoma City Public Schools.

Mehrens, W. A. (1984). "National Tests and Local Curriculum: Match or Mismatch?" *Educational Measurement: Issues and Practice*, 3, (3), 9-15.

Mehrens, W. A. and Kaminski, J. (1989). "Methods for Improving Standardized Test Scores: Fruitful, Fruitless or Fraudulent?" *Educational Measurement: Issues and Practices*, 8(1), 14-22.

Shepard, L. A. and Kreitzer, A. E. (1987). "The Texas Teacher Test." *Educational Researcher*, 16(6), 22-31.

Matching your curriculum and standardized tests

Jane Close Conoley, Buros Institute of Mental Measurements

What are the purposes of standardized tests?

Standardized tests can serve many different purposes for your school district. For example, standardized tests can

- help you evaluate your district by comparing it to a national sample of other districts,
- give you information about the success of various instructional programs in your district,
- give you an opportunity to judge how students in your district compare to their peers throughout the nation, and
- help you to diagnose particular strengths and weaknesses in individual students' learning.

Information related to each of these purposes can be found in the scores of most standardized achievement tests, but different tests are designed to accomplish some of these purposes better than others.

Most large-scale, commercially published achievement tests are designed to give you normative information. That is, test scores indicate how your district stands in relation to other districts. These tests are not the best choice for evaluating particular district programs or for diagnosing individual students' strengths or weaknesses. The items in these tests are based on a composite of information and skills that are taught at various grade levels all over the country. These tests do not reflect any particular district's curriculum. Therefore, they can never perfectly match your district's curriculum.

Although having a high degree of similarity between your district's curriculum and the items on a standardized test seems like the preferred situation, such a match limits the test's usefulness as a measure of your district's performance as compared to a national referent or to other identified referents. A high degree of similarity makes interpreting scores somewhat suspect. The norms are based on the performance of students from a wide variety of districts, not on the performance of students from districts whose curriculum is a very good match to the test items.

How do you match a test to your curriculum?

When you try to match a test with your district's curriculum, you should

- **Determine if you really have a good reason to match local objectives to test objectives.** The district-by-test-objective match is usually a source of random error and does not significantly affect a

district's relative standing among a national sample of districts.

- **Only if you have a good reason to match the test and the local objectives, determine the degree of match that is acceptable.** This is a complex decision because you can consider at least three levels of matching:

  1. the overall fit between the test and the curriculum,

  2. the fit of individual items to a content domain, and

  3. the impact of test specification on examinees' performance (Crocker, Miller, Franks, 1989).

  If you will use the test to make inferences or generalizations about broad categories of student learning, then perfect matches are unnecessary. For example, an 80% match of test and local objectives is acceptable.

  However, if you will use the test to evaluate a program or diagnose a student's strengths and weaknesses, then a high degree of match is necessary.

- **Prepare curriculum documents.** You must identify the learning objectives for each grade level.

  - Consider only those objectives that can be measured by multiple-choice items.

  - Note where literal interpretation of the objective is unnecessary. Similar items may be acceptable even if they don't fit the objective exactly.

- Cluster the objectives into reasonable domains. You may combine three or four specific objectives for a broader meaning.

- Code the objective clusters with unique numbers. For example, assign a grade number and letter to each course, followed by the specific cluster indicator (that is, 6M:31 is the sixth grade math, cluster number 31).

- **Prepare a table of specifications.** You'll need a team of content experts from your district to match test items with district objectives. Table 3 shows a sample table of specifications.

Table 3. A Sample Table of Specifications

Cluster code	Subtest and item numbers	Number of items
3M:017	M1: 23, 25, 29 M3: 17, 21, 30, 35	7
3M:018	M1: 25 M2: 19, 31, 33	4
3M:019	M1: 24, 26, 32, 25, 40	5

- **Prepare an item summary table.** List all the items in each subject and the cluster code to which each item refers. This table will give you a quick overview of the "tested but not taught" portion of any mismatch.

- **Prepare a summary table.** Indicate what portion of the objectives are matched by at least one item, at least two items, or at least three items. We show an example in Table 4.

Table 4. Objective Match Summary Table

Percent of cluster matched

Grade/ Subject	Number of Clusters	One or More Items	Two or More Items	Three or More Items
3 / Math	11	91%	73%	45%
Reading	14	95%	84%	79%

- **Evaluate the match.** The larger the number of objectives in each content domain, the lower the percent of match. With 14 reading objectives in grade 3 reading, it is reasonable to expect an 85% to 95% match. But if the same course is defined by 140 objectives, then the match (using the same items) will

seem less, unless the same item can be referenced several times.

If you give test publishers your district objectives clusters, they can do the matching for you. However, if they do, check the accuracy of their work by double-checking a sample of the clusters with the items the publisher says meet the objectives. We show you an example in Table 5.

Table 5. Accuracy of Publishers' Matching

Publisher	Clusters Checked	Number of Errors	Accuracy Index
A	50	4	92%
B	50	19	62%
C	50	15	70%
D	50	6	88%

What generalizations can be made from this discussion?

If you use the test to investigate the quality of education in your district, some may argue that a 60% match is better than a 90% match. If only formal teaching is tested, how can your district be sure that any other learning or growth has taken place?

On the other hand, if your district's goals are to evaluate its program with less concern about national norms, then a close fit is desirable.

References

Crocker, L.M., Miller, M.D., and Franks, E.A., (1989). "Quantitative Methods for Assessing the Fit Between Test and Curriculum." *Applied Measurement in Education,* 2, 179-194.

Using customized standardized tests

Paul L. Williams, CTB/McGraw-Hill

Over the next several years it is likely that you'll see a subtle but important change in the nature of standardized tests that are administered as part of your state and district testing programs. This change results from a desire to improve both the norm- and criterion-referenced interpretations of student, school, district, and state testing data. These interpretations can be improved by customizing the traditional norm-referenced test.

Norm-referenced tests are designed to give you both normative and objective information. Normative information may take the form of scale scores, percentile ranks, grade equivalents, normal curve equivalents, and stanines. Objective performance is usually reported as a percentage mastery score based on the objectives included on the norm-referenced test.

Normative scores allow you to compare individuals and groups with national performance levels, and objective scores allow you to make comparisons relative to specific objectives. Together, these scores allow you to plan programs for your school and district and instruction for individual students.

When used correctly, this information is invaluable for school administrators. However, several improvements can be made so that you can make even better programmatic and individual plans, such as

- reducing testing time,
- increasing the relevance of the test to the curriculum, and
- having greater confidence in the national comparative information.

These improvements are the goals of customized norm-referenced tests.

Several models for constructing customized norm-reference tests have been attempted, with some degree of success. A discussion of three models follows.

A model used in Texas

For the last few years, Texas has used a model state criterion-referenced test, which was statistically equated to a nationally normed norm-referenced test. Texas now administers the criterion-referenced test instead of the norm-referenced test and both norm-referenced and criterion-referenced scores are produced.

The advantages of this approach are reduced testing time and greater relevance to the Texas curriculum than could be obtained from using the norm-referenced test alone.

However, this approach has several disadvantages:

- Equating these two different tests will result in inaccurate norm-referenced scores because of differences in test difficulty and content between the norm-referenced and criterion- referenced tests. Criterion-referenced scores are unaffected by the equating.

- Instruction focused on the curriculum will likely increase both the criterion-referenced scores and, as a result, the equated norm-referenced scores. Although score increases on the criterion-referenced portion of the test may accurately reflect student learning in these restricted domains, this is not the case for the much broader norm-referenced domains.

  This is because instruction has been effectively focused on only a portion of the traits measured by norm-referenced tests, thus producing higher equated norm-referenced scores than would be expected if the original norm-referenced test or a proper sample of items from that test were administered.

  When this distortion happens, the norm-referenced scores produced from this model are called norm-invalid. That is, the customized test does not accurately reproduce the normative scores that would have resulted had the entire norm-referenced test been administered.

For a customized norm-referenced test to be fair, the scores must be norm-valid (Yen, Green, and Burket, 1987). Texas will leave this model in 1990 in favor of one

that may be more successful in producing scores that approach norm-validity.

A second model

A second model of a customized test is one in which state- or district-developed criterion-referenced items (consistent with local objectives) are combined with a complete norm-referenced test. Norm-referenced scores are generated from the complete norm-referenced test, while objective information is derived from a combination of norm-referenced and locally developed items.

This type of test reduces testing time because only one customized test is administered instead of both a norm-referenced and a criterion-referenced test. However, as with the Texas model that we discussed, norm invalidity may be a problem.

If instruction is carefully targeted at the objectives and a subset of the norm-referenced test items is used for reporting achievement by objective, then norm-invalidity could result because instruction influences only a portion of the trait measured by the norm-referenced test. In this case, the norm-referenced scores could be inflated by the targeted instruction, thus rendering them invalid.

A model used in Tennessee

Another model of a customized test was recently adopted by the State of Tennessee. The Tennessee model remedies the shortcomings of the first two models that we

described. This model uses approximately 40 items instead of a full-length test of 80 to 110 items for its norm-referenced module and a criterion-referenced module of state-developed items.

The norm-referenced module was specifically created so that it has proper statistical characteristics of reliability, adequate floors and ceilings, and articulation across test levels. Tennessee will use multiple test forms.

Items used for the norm-referenced portion are not intended to be used for objective scores, and the criterion-referenced items are not used as part of the norm-referenced scores.

Effective instruction targeted toward the state objectives will demonstrate student attainment of the state's objectives, and the norm-referenced portion will provide norm-valid scores. Thus, the Tennessee model reduces testing time and requires only one testing period rather than two. The objective scores will be useful for instructional planning and the norm-referenced scores can be used with confidence for national comparisons.

A note about norm-validity

As a school administrator, you should be concerned about the norm-validity of your district's test scores. During times of increased school, district, state, and national achievement (as we see now), critics may be quick to question the validity of your test results. Critics may point out that teachers are too familiar with the test items, that they teach actual test items, or that the scores

may not reflect true changes in achievement. Williams (1988) and Koretz (1988a, 1988b) have both presented a distinction between changes in test scores and changes in achievement.

Changes in test scores may result from a variety of instructional and administrative interventions, but changes in test scores may not reflect actual changes in achievement. Special coaching, inappropriate test preparation materials and methods, and narrowly targeted instruction may all increase test scores, but they do not necessarily lead to sustained and abiding increases in achievement.

Just as instruction must support test score changes that are genuine, test instruments must be designed and implemented so that if score increases occur, they represent a true change in achievement and are not the result of an inadequately designed customized testing program.

Unless a customized norm-referenced test produces norm-valid scores, you cannot provide test results that reflect true changes in achievement. Even with an optimally designed customized test, abuses can still result. But without a properly designed customized norm-referenced test, you cannot demonstrate that achievement, rather than just test scores, has improved.

Administrators at all levels must be able to tell the difference between norm-valid tests that allow actual achievement to be demonstrated and norm-invalid ones. When norm-valid test are used, you can report the test results with confidence.

If you have confidence in the test's quality, then test scores will accurately reflect meaningful changes in student achievement. Thus, you will be able to determine the effectiveness of your instructional program.

If you have a norm-valid test, you can show your constituents that changes in the test scores are real. When these changes represent increases, your community and staff can be satisfied the instructional program works in the areas the test measures. If the score changes represent a decrease, then the test results can help you identify areas that need additional instructional effort. In either case, the students win because instructional support is forthcoming.

Customized norm-referenced tests offer a viable alternative to both norm-referenced and criterion-referenced tests. One test, instead of two, is all that needs to be administered. Disruption in the schools is reduced, testing time is reduced, and instructional time is maximized. Alternate forms of customized norm-referenced tests can be used, minimizing criticisms of test familiarity and inappropriate test preparation activities. Teachers will be more likely to teach the complete curriculum, and increased achievement, rather than just increased scores, can result.

References

Koretz, D. (Summer, 1988a). "Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction?" *American Educator*, 8-15, 46-52.

Koretz, D. (1988b). Panel presentation, ECS Large-Scale Assessment Conference, Boulder, Colorado.

Williams, P.L. (1988). Panel presentation, ECS Large-Scale Assessment Conference, Boulder, Colorado.

Yen, W. M., Green, D. R., and Burket, G. R. (1987). "Valid Normative Information from Customized Achievement Tests." *Educational Measurement: Issues and Practice*, 6, 7-13.

Interpreting test scores for compensatory education students

Gary Echtemacht, Educational Testing Service

To follow the rules and regulations of compensatory education programs correctly, you must use objective measures when you select students for programs, assess their progress, and monitor the program's quality. Because you have this pressure to use standardized test scores, you should make sure that you use the tests correctly.

In this section, I point to four practices that administrators often mistakenly follow when they use test scores:

- using test scores alone to select students for programs,
- giving out-of-level tests,
- misinterpreting *grade-level*, and
- failing to differentiate the degree of error in individual and group scores.

Although these practices may not be widespread, they are serious.

Don't use test scores alone to select students for programs

Program regulations for Chapter 1 require that you select students by using objective measures. In addition, state departments of education sometimes impose other requirements -- for example, a program can serve only students who score below the 40th percentile rank or must serve all students who score below the 20th percentile rank.

These requirements often lead administrators to select students on the basis of test scores alone because

- the requirements are stated in terms of test scores, and
- when program monitors review programs, they appraise them in terms of state and federal regulations.

Nevertheless, you should not make a decision about an individual student by using a test score by itself. It is acceptable to use test scores to make decisions in a sequence of assessments, but it is unacceptable to use test scores by themselves in a sequence of one assessment. You are unfair to students if you simply say that all students who score below the 40th percentile rank are in the program and all who score above the 40th percentile rank are ineligible.

You must remember that test scores are neither completely reliable nor valid indicators of academic performance. For example, if students take an equivalent form of a test at different times, their scores will change somewhat. This unreliability is important for those whose

scores are near the cut-off score for selection because if you administer the same test a second time, some students who previously scored **below** a cut-off may score **above** the cut-off a second time.

Similarly, reading tests give you only general measures of reading ability. Some students may be good readers in certain content areas, yet they may score poorly on a given test because the reading passages in that test do not include the content areas they know.

Good programs select students by using several assessment tools, rather than just one. Although the regulations do not explicitly state other requirements, they do allow you to use additional assessment tools in selecting students. Ask your state director how you can best use other assessment tools, such as report card grades, results of other tests, and systematic teacher assessments obtained through questionnaires.

Some common methods for using multiple assessments are:

- selecting students who score below prescribed cut-offs on both your district's standardized test and another state-mandated test;

- using your district's standardized test to identify a pool of possible participants, then using either a teacher-completed questionnaire or report card grades to select students from the pool;

- using a systematic method for obtaining teachers' judgments about students' needs in order to identify a

pool of possible participants, then using a standardized test to select students from the pool; or

- using the standardized test to identify a pool of students, then creating a study team to select students from the pool and carefully documenting the study team's process.

Don't give out-of-level tests

Out-of-level testing occurs when you give a standardized test to students who are at a different grade level than the one for which the test is designed. In some cases, school officials use out-of-level tests in compensatory programs because those students are behind their peers or because in-level testing frustrates them. Administrators who follow this practice believe that somehow it is more valid to give those students tests designed for lower grade levels.

While out-of-level tests may be less frustrating to some students, the scores obtained from them are also less valid because

- the content for out-of-level tests does not represent the content taught in the classroom,
- the scale that test publishers use to link different test levels is loaded with error,
- there are no norms for out-of-level tests,

- scores obtained on tests of different difficulty are not comparable, and

- when obtained, out-of-level scores appear to be too low.

Although in-level test scores are more reliable in the middle than at the high- and low-score ranges, they are quite reliable in placing students at the high or low end of the scale. For example, with a reasonable degree of assurance, we can say that a student who scores at the 10th percentile rank is most likely a low-achieving student. What we are less sure about is whether the student is at the 10th percentile rank or the 15th percentile rank. Either way, we are reasonable in concluding that the student is low achieving.

You should use tests at the grade levels for which they are specified by the test publisher. Generally, the content of grade-level tests will represent what is taught in regular classrooms at the specified level.

If your compensatory program is good, it will be closely coordinated with instruction in the regular classroom. Because the purpose of compensatory education is to help students succeed in the regular classroom, using in-level tests will help you coordinate the two.

Understand the term "grade-level"

Generally, when school people say that certain students perform *at grade-level*, they mean that those students can learn material at about the same rate and quality as

others in the same class. The implication is that students who don't perform at grade-level have significantly more difficulty in class than their peers. Accordingly, when students are labeled as *working below grade-level*, the implication is that they may not have the aptitude, maturity, or interest to do the work that others in the same class are doing. Relatively few people are considered to be working below grade-level.

In contrast, in the testing arena *at grade-level* has a different meaning. When students score at grade-level, their scores are at the 50th percentile rank. It means that about half of their peers score higher and about half score lower. In testing, *at grade-level* does not relate to how well students perform in the classroom. Therefore, when you review students' scores, you must consider that, by definition, many students score below grade-level.

Historically, the term *grade-level* has been important in the politics of compensatory education. Advocates of compensatory education programs have always said that those programs were underfunded because many students who performed below grade-level did not receive program services. In this case, performing below grade-level was defined as scoring below the 50th percentile rank. While it is true that compensatory education may be underfunded and, I believe, is an important part of schooling, it is inappropriate to use the term grade-level in the testing-related sense and imply the general sense.

Because most people use the term grade-level in the general sense, you should either avoid using grade-equivalent test scores or develop a range of scores indicating satisfactory achievement in the classroom. You

may also think of average performance on a test as being between the 23rd and the 77th percentile rank.

Differentiate the degree of error in individual and group scores

Administrators tend to interpret differences in test scores in one of two ways. First, they may think that a difference of one or two percentile rank points is an important difference. Secondly, they may think that a difference of ten points shows that the test is unreliable. Few administrators can differentiate the degree of error in individual and group scores.

An individual test score is just that -- the score that an individual student receives on a test. A group score is the average of several individual scores. For example, the average score of third graders at Horace Mann Elementary School is a group score.

In general, individual scores have more error in them than group scores do. The error in an individual score is largely a function of the test's standard error that is described in the publisher's technical manual. For most of the tests given in elementary and secondary schools, the standard error is about 2.5 raw score points. This means that about 95% of the time, we would expect the scores for individual students to fall within a range of 10 raw score points. That is not particularly reassuring, but it is exactly why we need to use multiple measures for selecting students and why for most of the tests we use we should be a little skeptical of individual test scores and cautious in interpreting differences.

The error in group scores largely depends on the size of the group. Once you have a group of about 30 scores, the magnitude of the errors decreases. By the time you average all the scores for your school district, you can regard the results as accurate as long as there is not some systematic bias operating for most everyone in the district.

You can be confident of your interpretation when you consider score averages of large groups. For instance, when you consider a group of 55 scores, a score average change of one or two percentile rank points is an important change. If you consider averages based on fewer cases, you must be more cautious. You can be more or less confident of average scores depending on the number of scores. There is a definite hierarchy in the strength of your interpretations of test scores. Your interpretations are most sure when you consider district averages, followed in order by building averages, classroom averages, and finally individual students' scores.

Working with the press

Allan Hartman, Massachusetts State Department of Education

As an administrator, it is important that you establish good relations with the press. Because the press is often the vehicle through which the public gets information about what is happening in schools, it is through the press that schools are held accountable to the public.

When you work with the press, you should keep in mind two types of goals -- short-term and long-term.

- Your short-term goal is to **communicate specific information clearly and accurately so that it will be reported correctly.**

- Your long-term goal is to **build good relationships with the press to improve your future efforts of communicating information.**

Every time you deal with the press, be sure that what you do helps you meet these goals. Put simply, don't sacrifice a good long-term relationship with the press for the sake of today's story.

Here are guidelines that will help you meet these goals.

Presenting information to the press

- **Be clear.** Test scores can be complex to understand. Most people have trouble interpreting them fully and accurately. This is true of many other issues as well, such as school budgets.

  Don't expect reporters to plow through lengthy and complex reports to dig out important information. It's your school district's job to present the information clearly and directly. You're in a position to go through the information, extract the important facts, and present the them in a clear and straightforward way.

- **Be thorough.** Even though reporters may not want to wade through long reports to get the information they need, they often need backup data so that they can follow up on something interesting or document what they write. Always have complete and detailed information ready to give to them in a usable format, if they need it.

- **Be accurate.** If you give out information in such a way that it can be misinterpreted, it probably will be. Always make sure your information is correct. This is obviously essential, because you don't want to have to correct misinformation later. It's also essential for building a good long-term relationship with the press. Reporters need to know that they can trust your facts and figures.

- **Be honest.** If test scores in your district drop, don't try to hide it. First of all, you probably cannot. Secondly, you gain credibility by being honest. If you're willing to openly share what some may perceive as bad news, everyone will be more likely to believe you when you deliver good news.

- **Know your data.** You must be able to clearly and accurately answer questions that may arise. If you hesitate when you respond, people may wonder if you're trying to hide something. Also, if you're not sure of your information, you're more likely to become defensive. And, worst of all, if you're not thoroughly familiar with the data, you may give someone some wrong information.

- **Take the initiative.** Don't wait for reporters to knock on your door. You need to contact them. Remember, reporters don't want to miss a story that the community will find interesting, so let them know when there's information available. Also, reporters need to know that if a story unfolds in the future, you'll let them know about it.

- **Be timely.** Let the press know when certain kinds of information will be available. Also, know and be aware of the time pressures that reporters have, such as deadlines, printing dates for local papers, and taping times for radio shows. If you want reporters to understand your problems, you have to be fair to them by understanding their problems and constraints.

If you often find yourself saying, "But they didn't know the situation here," when you read a story about your district, you need to do something about that. You should take the following three steps to ensure that people have the information they need.

Stage 1: Background

1. **Prepare a background kit for yourself,** using the following checklist. You'll probably want to update this information periodically, because it may change over time. If you have this information compiled ahead of time, you'll be ready to contact the appropriate people when it's necessary.

- ☐ Names, addresses, and phone numbers of all local and regional newspapers
- ☐ Names and phone numbers of any reporters who specifically cover education or civic affairs.
- ☐ Deadlines and printing days for newspapers
- ☐ Names, addresses, and phone numbers of all local or regional radio and television stations
- ☐ Information about which shows are daily and which are weekly, including taping times and times of live shows

2. **Prepare a background kit for the press**, using this checklist. Remember, you'll have to update this information every year to keep it accurate.

- ☐ Names, addresses, and phone numbers of top school administrators
- ☐ Names and ways of contacting School Board members
- ☐ Information about your district, including
  - ☐ enrollment
  - ☐ number of teachers
  - ☐ size of your annual budget
  - ☐ number of students in special education programs
  - ☐ proportions of students in various racial or ethnic groups
  - ☐ number of male and female students
  - ☐ number of students at each grade level
  - ☐ names, addresses, enrollments, and grade levels for each school in your district

- ☐ Information about your testing program, including
  - ☐ names of all standardized tests that your district administers
  - ☐ approximate annual testing period
  - ☐ approximate dates for reporting results
  - ☐ grades that are tested
  - ☐ for Basic Skills tests, the score required to meet your district's standards
  - ☐ calendar of events, including
    - annual testing dates
    - annual dates when test results are available
    - major dates in the budget preparation process
    - cycle for School Board elections
    - annual school events (for example, fairs, graduation)

3. Once you have prepared and reproduced your background kit, **give the information to the press**. The presentation of this kit may give you an opportunity to meet local reporters and broadcasters. For instance, you might meet with members of the

press one at a time or you might conduct a tour of your schools and host a general-information, get-acquainted session to distribute your packet. When you invite reporters to this background briefing, however, be sure to let them know that there's no "hot story" happening. Also, take this chance to find out what other kinds of information they need as background on your district.

Stage 2: Getting the news out

Let's say you've just gotten your district's test results and you're about to make a formal presentation at next week's School Board meeting. It's time to talk to the press. These three steps will help you get the information to the right people at the right time.

1. **Write a press release.** Make sure your press release includes the following information. (*See the end of this section for a sample press release.)*

   - **What are the test results?** How does your district compare with other districts in your state? What are the percentile scores for each grade? Where do the percentile ranks fall in relation to the comparison score band?

   - **How do the test results compare with last year's results?** Is the trend higher or lower?

   - **If the trend is higher or lower, what is the reason?** Budget cutbacks? A new reading program? Shorter instructional periods? A redesigned curriculum?

- **What are you going to do with the test results?** Create a new program of some kind? Reallocate funds? Convene a new task force? If you're going to start a new task force, when will the public see something materialize?

- **How do you feel about the test results?** If you're pleased, say so. If not, it's good to express your concern.

2. **Call reporters to let them know when and where you will release the information.** You can give them the press release ahead of time, which would give them the date and time when they can publish the news.

   For instance, if the School Board will meet on Monday night, you can call reporters on Friday, give them the press release, and answer their questions. They will not release the results until the time specified on the press release, but having the press release ahead of time allows them to gain a better understanding of the information you will present and to ask questions. This early presentation of information helps both your district and the press.

3. **Give the press release to all reporters at the same time.** Don't play favorites -- you'll be sorry later, even if the favored reporter writes a great story this time.

   If the number of reporters is small, you might call them and ask if they want to pick up a copy of the

press release at a certain time when someone will be available to discuss it and answer their questions. Or you may want to mail the press release to all of them at the same time (if it is ready early enough to allow for mail delivery).

Stage 3: Follow-up

The test results are out and the press has presented the information clearly and accurately -- because you presented it to them clearly and accurately. But your job is not finished!

Keep in contact with reporters throughout the year. Tell them what is being done in your district as a result of the test scores. For instance,

- Are you planning to make any changes in the curriculum?
- What effect will the projected budget have on your program?
- Are you planning any new programs or any special events?
- Have any students or teachers won an award?

Let the press know what is happening in your district as it is happening. All of these events don't make timely news stories and they may not be reported. However, if they are reported, your community will be better informed. If they are not reported, at least the members

of the press will build up knowledge about your system; that will help them report the next news story with greater understanding.

See the following pages for a sample news release and news information memorandum.

102

Massachusetts Educational Assessment Program

Sample News Release and News Information Memorandum

The sample news release and news information memorandum on the following pages may help you communicate information about statewide assessment results. Both the news release and the memorandum report on different situations in a fictitious school district called Seaview. The purpose of these samples is to illustrate different approaches you could take to getting the news about test results out to the public.

The news release is written in journalistic style; it contains most of the information the media will need to report adequately on your district's assessment results. You may consider preparing a news information memorandum if you are unfamiliar with journalistic writing. The memo is simply an organized summary of all the pertinent details with the same information as the news release, but not written in paragraph form.

The content of the samples is neither definitive nor comprehensive. You would want to modify or add to the content, depending on your results and the emphasis that makes sense. At times, you may want to give the media a release or memorandum along with a report that you will make at a regular School Board meeting. The Board report would most likely contain details about individual school results.

Regardless of how you release the assessment results, most reporters would want to know at least the following information:

- What are the assessment results?
- How do the results compare with the earlier assessment? the state?
- If the results are higher or lower, what is the reason?
- What are you going to do with the assessment results?
- How do you feel about the assessment results?

Once you cover this basic information, they may also want to know about

- unusually high or low performance by grade level, subject area, or individual schools;
- numbers and percentages of students taking the assessment;
- purpose and design of your assessment program;
- background information about the students and factors composing the comparison-score band; and
- questionnaire results.

Sample News Release

Seaview School District **Release Date:** 11:00AM, Nov. 16, 1988
James J. Jones
Superintendent of Schools
(518) 777-7777

Seaview Test Scores Remain Stable

Background on testing: who, what, when, and highlights

Dr. James Jones, Superintendent of Schools for the Seaview School District, today announced the results from last spring's statewide assessment testing of all students in grades 4, 8, and 12. The results, reported for all schools as well as the district in reading, mathematics, science, and social studies, were recently released by the State Education Department as part of a mandated statewide testing program.

"The results show," stated Dr. Jones, "that with few exceptions, Seaview schools score at about the state average, which is similar to their performance at grades 3, 7, and 11 in 1986. Scores in most areas were somewhat higher at the elementary level and the strongest performance was in science compared to other subject areas."

The average statewide scores for both 1986 and 1988 were set at 1300 with a range of 1000 to 1600. Differences of less than 50 points are not considered by the State Education Department to be significant. The results for both years show:

Major results

Grade	Reading	Math	Science	Social Studies
----------1988----------				
4	1310	1320	1340	1310
8	1260	1290	1310	1220
12	1240	1290	1300	1260
----------1986----------				
3	1320	1290	1320	not assessed
7	1280	1300	1320	not assessed
11	1240	1280	1290	not assessed

Seaview schools performed somewhat higher than similar kinds of communities, especially in science. The district's scores were about the same as for districts enrolling students coming from similar backgrounds.

More about the tests

The state's assessment program is intended to survey a broad range of student achievement, including basic as well as higher order skills. To provide for this broad range of coverage, the Program administered over 3,000 test questions. To minimize testing time these questions were distributed across many different test forms, and each student completed only one form. Therefore, while highly reliable building and district results are reported, the program is not designed to yield student results.

What the district plans to do with the results

"While the assessment results are useful in informing us of how well our schools are doing in a comparative sense," Dr. Jones noted, "the more important use is in helping direct ways we may improve our school programs." According to Dr. Jones, a district-wide curriculum committee has been reviewing the results of both assessments and will be soon making recommendations. In addition, at the middle school level, reading personnel are currently reviewing the textbooks and supplementary reading materials to see if changes are warranted.

More details on test results

The test results indicated both strengths and weaknesses in each subject area. Strengths in mathematics across all grade levels included numbers and numeration, recognition of plane and solid figures and estimation skills. Relative weaknesses were mainly in the areas of fractions/decimals, some problem solving skills and probability and statistics. In science, schools demonstrated strengths in the life sciences and weaknesses in the physical sciences. Methods of scientific inquiry were relatively weak at the 12th grade level. At 4th grade, major strengths in reading were in literal comprehension and study skills. Weaknesses in reading at 12th grade included analyzing texts and some areas of study skills. At all levels in social studies schools showed strengths in areas of history but weaknesses in geography and map and research skills.

Limits and caveats

"Like all tests," Dr. Jones commented, "the assessment tests do not necessarily measure what is taught and learned at particular grade levels. Aspects of geometry, for example, that were tested at the 4th grade level are not introduced until the later grades."

Further, the superintendent noted, the tests do not measure all that is taught at particular grade levels. In the social studies, for example, there are aspects of government that are taught that were not included in the assessment.

Conclusion or summary statement

"Overall," concluded Dr. Jones, "we believe that the results will be informative and useful to us in building stronger instructional programs in our district. We expect that the improvement efforts, now underway, will be reflected in future assessments."

Sample News Information Memorandum

Seaview School District
500 Main Street
Seaview, Massachusetts 30000

For Release: November 16, 1988 - 11:00 a.m.

For more information:

Dr. James J. Jones
Superintendent of Schools
(518) 777-7777

MASSACHUSETTS EDUCATIONAL ASSESSMENT PROGRAM

News Information Memorandum on the Seaview School District's

1988 Scores on the Massachusetts Educational Assessment

1. **About the assessment program**: In April, 1988, 3200 Seaview 4th, 8th and 12th grade students were administered the state's mandated assessment tests in reading, mathematics, science and social studies. Results from the testing were recently released by the State Department of Education. Comparable testing of 3rd, 7th and 11th grade students was reported in 1986.

   The assessment tests are broad ranging achievement tests in each subject area and consist of over 3,000 test items. To minimize testing time the test questions were distributed over many different test forms with each student only completing

one form of the test. Therefore, only building and district reports are reported.

2. **Who was tested**: About 95% of all eligible 4th, 8th and 12th grade students took the tests, which is slightly above the state average. Most students in bilingual classes were exempted as were special needs students whose parents so requested.

3. **Assessment results highlights**: District performance was significantly above the state average on 4th grade reading and mathematics and 8th grade science; at the state average on 4th grade science and social studies, 8th grade reading, mathematics and social studies and 12th grade social studies and science and below the state average on 12th grade reading and mathematics. Overall, assessment scores were generally higher at the early grades than at the secondary grades during both assessments. Scores in science were most often higher than in other subject areas.

4. **Assessment Scores**: For each subject area the range of possible scores is 1000-1600, with a statewide average of 1300. Only differences of more than 50 points are considered meaningful by the State Department of Education. Scores below are for both 1988 and 1986 even though different grades were assessed in these years.

1988

Grade	Reading	Mathematics	Science	Social Studies
4	1390	1380	1340	1330
8	1320	1290	1360	1300
12	1240	1240	1340	1290

1986

Grade	Reading	Mathematics	Science	Social Studies
3	1340	1330	1320	not assessed
7	1300	1310	1370	not assessed
11	1250	1220	1330	not assessed

5. **Other Assessment Information:** Scores reported by the state also show how a district's scores compare to those of districts with students coming from similar backgrounds. Background factors considered are parental education, family language and the socio-economic conditions of the community. Seaview scores were above those of similar districts in 4th grade reading and comparable to similar districts in all other instances.

   Other information collected during the assessment showed Seaview students (a) demonstrating more interest in science than the statewide average, (b) spending more time on homework than the statewide average, and (c) writing more reports and papers in school than the average of their statewide counterparts.

6. **Actions Being Taken**: Several steps are being taken related to the new assessment scores:

   English and mathematics committees at the high school are currently reviewing program offerings and graduation requirements and will make a report in early January.

   A district-wide in-service program on critical reasoning skills is scheduled for February.

   Recommendations for a new mathematics resource center will be presented to the school committee by Dr. Jones in the late winter.

7. **Other Background Information**: The state report on background factors showed that the level of education of Seaview's parents was above the states average and that for 91% of parents the predominate language in the home was English. In addition, a larger percentage of special needs students participated in the assessment than the percentage statewide and fewer students were absent from the testing than the statewide average.

Appendices

- Finding information about standardized tests
- Organizations that provide test information
- Putting test scores in perspective: Communicating a complete report card for your school
- Major achievement tests and their characteristics
- Names and addresses of major test publishers
- A glossary of testing terms

Finding Information About Standardized Tests

Lawrence M. Rudner and Kathryn Dorko, ERIC/TM

Finding the right standardized achievement or aptitude test can be quite difficult. You need to identify a variety of potentially useful tests, collect and review technical materials, and identify and evaluate the practical considerations of using these tests.

This section is designed to help you with the first step -- identifying useful standardized tests. In it, we describe

- books that describe available tests,
- test reviews,
- online information retrieval systems, and
- other sources for testing information.

The printed sources are available in most academic libraries. They only contain brief information about individual tests; they do not contain copies of the tests themselves. You'll probably want to contact test publishers for more detailed information.

Books that describe available tests

The following books have basic, non-evaluative information about a wide range of available tests. All include statements about intended audience, publication date, scoring, author, costs, and publisher.

- Mitchell, James V. Jr. (ed.), *Tests in Print III (TIP III): An Index to Tests, Test Reviews, and the Literature on Specific Tests.* Buros Institute of Mental Measurements, University of Nebraska Press, 901 North 17th Street, Lincoln, Nebraska 68588-0520, (402) 472-3581, 1983, 714 pages.

  Tests in Print describes more than 2,400 published tests. It also contains more than 16,000 references about specific tests, a cumulative name index for each test that covers all references in *TIP III,* a directory of test publishers with all the tests of each publisher listed, a title index that covers all tests in print and all out-of-print tests once listed in *Mental Measurements Yearbooks (MMY)*, a name index to authors of more than 70,000 documents (tests, reviews, excerpts, and references) in the nine *MMYs* and *TIP III*, a scanning index for quickly finding tests that are designed for particular populations, and serves as an index to the MMY series in general.

- Keyser, Daniel J., and Sweetland, Richard C. (eds.), *Tests: A Comprehensive Reference for Assessment in Psychology, Education, and Business* (2nd ed.). Test Corporation of America, 4050 Pennsylvania, Suite 310,

Kansas City, Missouri 64112, (816) 756-1490, 1986, 1,296 pages.

This book concisely describes more than 3,100 published tests in a "quick-scanning, easy-to-read" format. It gives a brief description and information about the population targeted by the test, the purpose, and administrative and publication information.

- *The Educational Testing Service Test Collection Catalog, Volume I: Achievement Tests and Measurement Devices.* Oryx Press, 2214 North Central Avenue, Phoenix, Arizona 85004-1483, (800) 457-6799, 1986, 296 pages.

This catalog gives information about more than 2,000 achievement tests in the ETS Test Collection. It indexes tests by author, title, and subject category.

- Krug, Samuel E. (ed.), *Psychware Sourcebook 1988-1989.* Test Corporation of America, 1988, 640 pages.

This book describes 450 computer-based products used in psychology, education, and business. Most products go beyond simple test scoring and involve administration and report generation. The book has five indices: Test Title, Product Category, Product Application, Service, and Supplier.

- Pletcher, Barbara P., Locks, Nancy A., Reynolds, Dorothy F., and Sisson, Bonnie G. *A Guide to Assessment Instruments for Limited English Speaking Students*. Santilla Publishing Company, New York. Out-of-print. Available through ERIC Document Reproduction Service, 3900 Wheeler Avenue, Alexandria, Virginia 22304, (800) 227-3742, TM 011 805, 1977, 223 pages.

  While somewhat dated, this reference gives you leads to assessment instruments for native speakers of Chinese, French, Italian, Navajo, Portuguese, Spanish, and Tagalog. The instruments listed in this guide were designed for use with students in K-6 and were normed with students in the U.S. Descriptive, technical, cultural, and linguistic information is given for about 400 tests.

Test reviews

Several major books give in-depth, candid reviews of available tests. The best-known books are:

- Mitchell, James V. Jr. (ed.) *The Tenth Mental Measurement Yearbook*. Buros Institute of Mental Measurements, 1989, 1,014 pages.

  The *Yearbooks*, published periodically since 1932, are a comprehensive source of factual and evaluative information about commercially available tests. *The Tenth Mental Measurement Yearbook*, contains information about 396 tests and includes 569 reviews by 303 different authors. In addition to descriptive

information and test reviews, this book has bibliographic references to studies and articles about specific instruments, and a current directory of test publishers.

- Keyser, Daniel J., and Sweetland, Richard C. (eds.), *Test Critiques*. Test Corporation of America, Volume I, 1985, 800 pages; Volume II, 1985, 872 pages; Volume III, 1985, 784 pages; Volume IV, 1986, 768 pages; Volume V, 1986, 608 pages; Volume VI, 1987, 712 pages.

  Test Critiques emphasizes the practical aspects of test administration. Each review in this series has an introduction, practical applications, technical aspects, and an overall critique of the test.

Online information retrieval systems

Identifying and searching test information can be done quickly and efficiently through the online database system managed by Bibliographic Retrieval Services (BRS), 1200 Route 7, Lantham, New York, 12110, (800) 468-0908.

BRS provides sophisticated search routines and access to databases that contain test information. You or your librarian can search by test title, parts of a title, subject, purpose, availability, grade level, or any combination of these and other descriptors. The following testing databases are available:

- *The Educational Testing Service File (ETSF)*

This is an online index to the tests contained in the Educational Testing Service (ETS) Test Collection. Developed to support the work of ETS test development staff, the ETS Test Collection has more than 14,000 commercial and unpublished tests. More than 8,000 tests that are currently available are in the ETSF.

- *Mental Measurements Yearbook Database (MMYD)*

This is an online index to 1,400 tests and reviews covered in the *Mental Measurement Yearbooks*. Although considerably smaller than the ETSF database, the *MMYD* has more detailed information about each test and more information that can be searched.

Other sources of testing information

Other sources for testing information are described in:

- Fabiano, Emily, and O'Brien, Nancy. *Testing Information Sources for Educators*. ERIC Clearinghouse on Tests, Measurement and Evaluation, American Institutes for Research, 3333 K Street, NW, Suite 300, Washington, DC 20007, (202) 342-5060, Report TME-94, 1987, 61 pages.

This is a guide to more than 150 books, journals, indexes, and computer-based services and organizations that provide information about student assessment. It also includes a subject index.

- Crosby-Muilenburg, Corryn. *Psychological and Educational Tests: A Selective Annotated Guide*. ERIC Document Reproduction Service (TM 011 545), 1988, 35 pages.

  Developed as a guide to the extensive measurement resources available to patrons of the Humbolt State University Library (Arcata, CA), this report identifies a wide range of books, reports, and journals about tests. It includes an extensive listing of references within specific disciplines, such as special education, counseling, and early childhood.

Organizations that provide test information

Ronald T.C. Boyd, ERIC/TM

The following organizations provide information, services, or publications related to testing. Most of these organizations provide print material. Some also provide testing services, access to test collections, or speakers on issues that involve testing.

American College Testing Program (ACT)
Box 168
Iowa City, IA 52243
(319) 337-1000

ACT emphasizes career and educational assessment. ACT provides services for students who seek college admissions, advising, financial aid, career planning, continuing education, and professional certification. ACT also administers the ACT college admissions test. A free catalog is available.

Association for Measurement and Evaluation in Counseling and Development (AMECD)
5999 Stevenson Avenue
Alexandria, VA 22304
(703) 823-9800

121

AMECD serves people who plan, administer, and conduct testing programs. It identifies problems in applying tests, promotes research in testing, provides test scoring services, interprets test results, and develops evaluation instruments. A free catalog is available.

Buros Institute of Mental Measurements
135 Bancroft Hall
University of Nebraska
Lincoln, NE 68588-0348
(402) 472-6203

The institute publishes the Mental Measurement Yearbook, which contains factual and evaluative information on recently released tests. The institute also maintains a comprehensive collection of commercially available tests, a historical library of tests, and books related to testing.

Center for Research on Evaluations, Standards, and Student Testing (CRESST)
School of Education
University of California, Los Angeles
145 Moore Hall
Los Angeles, CA 90024-1521
(213) 825-4711

CRESST seeks to improve education through systematic evaluation. Research focuses on three aspects of testing: testing for improved learning; systems for evaluation and improving educational quality; the impact of testing on educational standards, policies, and practices. A free catalog is available.

Educational Testing Service (ETS)
Rosedale Road
Princeton, NJ 08541
(609) 921-9000

An educational testing and research organization that provides test related services for schools, colleges, and government agencies. ETS also administers testing programs for agencies such as the College Entrance Examination Board's College Level Examination Program (CLEP), the Graduate Record Examination (GRE), and the Scholastic Aptitude Test (SAT). ETS maintains a library with 15,000 titles and a large collection of current and out-of-print tests. A free catalog is available.

ERIC Clearinghouse on Tests, Measurement, and Evaluation (ERIC/TME)
American Institutes for Research
3333 K Street, NW, Suite 200
Washington, DC 20007
(202) 342-5060

ERIC/TME acquires, selects, and abstracts documents on testing and evaluation for the Educational Research Information Center system. It also provides a variety of information products including books, annotated references, and information flyers. A free catalog is available.

Evaluation Assistance Center (EAC)--East
Georgetown University
1916 Wilson Blvd., Suite 302
Arlington, Virginia 22201
(800) 626-5443

Evaluation Assistance Center (EAC)--West
College of Education
University of New Mexico
Albuquerque, NM 87131
(800) 247-4269

These two regional centers collect and summarize published tests that are designed for limited English proficient (LEP) students. The EACs also provide technical assistance to educational programs that serve LEP students.

International Reading Association
P. O. Box 8139
800 Barksdale Rd.
Newark, DE 19714-8139
(302) 731-1600

The association offers many publications on reading, including attitudes toward reading and measuring reading performance. Most publications are under 50 pages.

Lawyers' Committee for Civil Rights Under Law
1400 Eye Street, NW, Suite 400
Washington, DC 20005
(202) 371-1212

The committee researches legal issues that relate to individual civil rights, including civil rights pertaining to federal and state testing. It investigates questions of discrimination in employment tests and teacher performance tests.

National Assessment of Educational Progress (NAEP)
P.O. Box 6710
Princeton, NJ 08541
(609) 734-1624

NAEP conducts national surveys of basic skills in reading, writing, mathematics, science, literature, art, music, social studies, computer skills, citizenship, and career development. A free catalog is available.

National Association of Test Directors
341 S. Bellefield Ave.
Pittsburg, PA 15213
(412) 622-3940

The association assists people who are responsible for developing, administering, and interpreting tests in city and county public school systems. It provides speakers on subjects related to testing and publishes two newsletters, occasional papers, and a yearbook for its members.

National Center for Fair & Open Testing (FairTest)
P.O. Box 1272
Cambridge, MA 02238
(617) 864-4810

A research and public interest group that monitors the educational testing industry. FairTest publishes a newsletter, monographs on testing, and journal reprints that focus on test topics. FairTest also provides testing experts who give workshops, presentations, and testimony in court cases that involve testing issues.

National Council on Measurement in Education (NCME)
1230 17th street, NW
Washington, DC 20036
(202) 223-9318

A professional council of test publishers, measurement specialists, and educators who measure human abilities, personality characteristics, and educational achievement. They also publish a quarterly technical journal and a quarterly journal dedicated to issues and practices.

Northwest Regional Educational Laboratory
101 S. W. Main Street
Portland, OR 97204
(503) 275-9500; (800) 547-6339

The laboratory operates the Center for Applied Performance Testing and the Test Center. The Center for Applied Performance Testing conducts research on testing and develops training material for improving the quality of tests. The Test Center is a collection of tests and other assessment tools. A free catalog is available.

Test Information Center, Chapter 1 Technical Assistance Center
Educational Testing Service
1560 Sherman Avenue, Suite 300
Evanston, IL 60201
(312) 869-7700

TIC/TAC collects and reviews published achievement tests for Chapter I programs. Their publications are available through the regional Chapter 1 Assistance Centers and ERIC/TM.

Putting test scores in perspective: communicating a complete report card for your schools

M. Kevin Matter, Cherry Creek Schools, Colorado

Research and Evaluation staff often receive numerous telephone calls from people who want to know which school is the *best* in the district -- or which one has the highest test scores. Invariably, these people equate the best with highest scores. What most of these people want is a school that will challenge their children intellectually, emotionally, physically, and socially. And, they believe that if their children are with those students who have the highest scores -- that is, the *best*, by their standards -- then their children will be better.

In light of this expectation, you should present the most complete and impartial picture of your schools to your students, their parents, and the community. However, just sending a report on test scores, even though the scores may be above the state, national, or district average, is insufficient because you must communicate about the total educational program in your district.

But what else can you use to measure your district's performance in preparing students for life?

Additional measures are available. Some require effort to collect and organize; others are easily gathered and summarized if you outline and implement a process. By collecting the information on some of these other important indicators of your district's work, you may be better able to interpret and use standardized test data to improve your educational program. More important, your community will learn more about what education is as a profession.

Factors to consider when you communicate about your schools

Attendance

- Absences: students
- Absences: staff
- Tardies: students
- Tardies: staff
- Percent of students and staff with perfect attendance or fewer than *x* absences
- Student participation in before/after school programs
- Parent participation in PTO meetings, back-to-school nights, special programs

- Total enrollment
- Graduation rate
- Dropout rate

Non-student participation

- PTO membership
- Number and types of parent volunteers
- Number and types of special programs, fundraisers, etc. (for example, RIF, Jr. Great Books)

Diversity

- Student population by sex
- Staff population by sex
- Student population by ethnic group
- Staff population by ethnic group
- Percent of students who receive special educational help
- Percent of staff by responsibility (for example, regular classroom teacher, resource room teacher, clerical staff, support staff)

- Percent of students eligible for free or reduced-price meals
- Percent of students who have a home language other than English
- Percent of students who are eligible for Chapter 1 services

Stability

- Percent of students who are new to school/district
- Percent of staff who are new to school/district

Staff experience

- Average number of years of experience in the district
- Average number of years of experience in the school
- Average number of years of experience in education

Staff development

- In-service programs for teachers
- Peer coaching or teaching programs
- Collaborative programs between business and industry and the district

130

- Collaborative programs between colleges and universities and the district
- Coursework or training taken by staff during the year and during the summer
- Staff and school grants

Programs for students

- Study skills
- Counseling services (including vocational, post-secondary)
- Dropout prevention
- Students at risk
- Dropout recovery
- Preschool
- Peer or cross-age tutoring
- Community: Big Brother, Big Sister, Scouts, 4H
- Summer school
- Critical thinking, creative problem-solving

Achievement

- Students' performance after they leave: Feedback from middle and high schools on how well students are prepared in relation to students who attend other elementary or middle schools in the area
- Special projects by teachers, parents, and staff
- Faculty, staff, student awards, presentations, publications, honors
- Percent of staff with advanced degrees
- Standardized test scores
- Local assessment results
- Previous year's test scores
- Test scores for cohorts (following the same group of students throughout their school careers)
- Distributions of test scores (percent of students who scored above the 75th percentile, below the 25th percentile, etc.)
- Number of books checked out of the library per student
- Accreditation
- Number of National Merit Scholarship: qualifiers, semi-finalists, finalists

- Student retention rate and number
- Excellence rewards
- Average number of high school out-of-class accomplishments
- College entrance examination (SAT, ACT) averages

Environment

- Number of incidence of student vandalism
- Number of fights between students
- Types and numbers of disciplinary actions against students
- Number of fights between staff (just kidding -- wanted to see if you got this far)
- Special services at the school for:

  physically handicapped students
  emotionally handicapped students
  socially handicapped students
  mentally handicapped students
  academically low-achieving students
  academically gifted students
  talented students (academic and nonacademic)
  students from low-income families
  students dominant in a language other than English
  students with learning disabilities

students with behavioral problems

- Extracurricular activities at the school for students:

  instrumental music
  vocal music
  sports
  clubs
  interest groups

- Number of hardbound library books per student
- Number of computer systems per *x* students
- Number of students who need (use) extended day services, both before and after school
- Average class size
- Student, teacher, administrator, staff support ratio
- Length of school day
- Length of school year
- Length of class periods
- Units required for graduation or advancement
- Average number of units taken in various subject areas
- Percent of students taking a foreign language (or other subjects)

- Average amount of homework required (by subject area)
- Percent of the school day of actual academic learning time

Fiscal

- Average teacher, administrator, staff support salary
- Expenditures per pupil
- Decentralized budget

Major achievement tests and their characteristics

Northwest Regional Education Laboratory

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Basic Achievement Skills Individual Screener, 1983 Psychological Corporation	N/A	Oct 12 (1-12)	**Wide Range Test** 1-12	Individually administered basic skills test for use in screening students, preparing IEPs, and placing transfer students.
California Achievement Tests, 1985-86 Forms E & F CTB/McGraw-Hill	Scale score	Oct 22 (K-12) May 2 (K-12)	**10**: K.0-K.9; **11**: K.6-1.9; **12**: 1.6-2.9; **13**: 2.6-3.9; **14**: 3.6-4.9; **15**: 4.6-5.9; **16**: 5.6-6.9; **17**: 6.6-7.9; **18**: 7.6-8.9; **19**: 8.6-10.9; **20**: 10.6-12.9	Level 10 measures readiness rather than achievement. It is not linked to other levels. Quarter month norms interpolated 3-6 weeks from the empirical week of standardization are available.
Comprehensive Assessment Program Achievement Series, 1980 Forms A & B American Testronics, Inc.	Equal interval score	Oct 15 (K-12) Apr 23 (K-12)	**4**: Pre K-K.5; **5**: K.0-1.5; **6**: 1.9-2.5; **7**: 2.0-3.5; **8**: 3.0-4.5; **9**: 4.0-5.5; **10**: 5.0-6.5; **11**: 6.0-7.5; **12**: 7.0-9.5; **13**: 9.0-11.5; **14**: 11.0-12.9	The publisher recommends testing no more than two levels out of level for functional level testing. Equated to the NTBS (1985).

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Comprehensive Tests of Basic Skills, 1981-82 Forms U & V CTB/McGraw-Hill	Scale score	Oct 14 (K-12) Apr 29 (K-12)	**A**: K.0-K.9; **B**: K.6-1.6; **C**: 1.0-1.9; **D**: 1.6-2.9; **E**: 2.6-3.9; **F**: 3.6-4.9; **G**: 4.6-6.9; **H**: 6.6-8.9; **J**: 8.6-12.9; **K**: 11.0-12.9 (see comments on level K)	The publisher provides Quarter Month norms interpolated 2-6 weeks from the empirical norm date. The publisher recommends testing no more than two levels out of level for functional level testing. Level K is more appropriate for testing students in college preparatory programs.
Degrees of Reading Power, 1979-83 Forms PA, PB & CP The College Board	DRP units	Nov 1 (4-12) May 14 (3-12)	**PA/PB-8**: 3.5-5.9; **PA/PB-6**: 5.0-7.9; **PA/PB-4**: 7.0-9.9; **PA/PB-2**: 9.0-12.9; **CP/1B**: 12.0-14.9; **CP/1A**: 12.0-14.9	DRP Units comprise a continuous scale across all test levels. Grades suggested for test levels are approximate. Selection will depend on abilities of students in the test group.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
DMI Mathematics Systems, 1983 CTB/McGraw-Hill	Scale score	See comments	**A**: K.6-1.5; **B**: 1.6-2.5; **C**: 2.6-3.5; **D**: 3.6-4.5; **E**: 4.6-5.5; **F**: 5.6-6.5; **G**: 6.6-8.9+	Percentiles and NCEs for the total score are linked to the CTBS/U (1981), the CAT/C (1977), and the CAT/E (1985) and are available through the publisher's scoring service. Norm-referenced scores can be provided for CAT or CTBS empirical or projected norm dates when the Instructional Objective Inventory is used. Norm-referenced scores are not available for hand scored tests.
Gates-MacGinitie Reading Tests, 1978 Forms 1, 2, & 3 Riverside Publishing Co.	Extended scale score	Oct 15 (1-10) Feb 15 (1) May 15 (1-12)	**Basic R**: 1.0-1.9; **A**: 1.5-1.9; **B**: 2; **C**: 3; **D**: 4-6; **E**: 7-9; **F**: 10-12	Supplementary out of level norms tables are available. The Basic R level should not be used out of level beyond grade 2.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Individualized Criterion Referenced Tests, 1979 Forms A & B Educational Development Corp.	Achievement scale	Oct 5 (1-8) May 1 (1-8)	**Primary I**: 1.0-1.9 **Primary II**: 2.0-2.9 **Elementary I**: 3.0-4.9 **Elementary II**: 5.0-6.9 **Intermediate**: 7.0-8.9	The Achievement Scale "links" nearly all of the ICRT (Form A) test booklets to a continuous Rasch scale. Criterion-referenced scores provided. Microcomputer software scoring program.
Iowa Tests of Basic Skills, 1985-86 Forms G & H Riverside Publishing Co.	Developmental standard score	Oct 31 (K-9) Apr 30 (K-9)	**5**: K.1-1.5; **6**: K.8-1.9; **7**: 1.7-2.6; **8**: 2.7-3.5; **9**: 3; **10**: 4; **11**: 5; **12**: 6; **13**: 7; **14**: 8-9; **TAP 15**: 9; **16**: 10; **17**: 11; **18**: 12	Expanded standard scores on the ITBS, Form G, are continuous with those of TAP, Form G, on some tests.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Iowa Tests of Basic Skills, 1978-82 Forms 7 & 8 **Tests of Achievement and Proficiency, 1978** Form T Riverside Publishing Co.	Standard score	Oct 28 (K-3) May 2 (K-3) Oct 30 (3-9) Apr 28 (3-9) Oct 29 (9-12) Apr 21 (9-12)	**5**: K.1-1.5; **6**: K.8-1.9; **7**: 1.7-2.6; **8**: 2.7-3.5; **9**: 3; **10**: 4; **11**: 5; **12**: 6; **13**: 7; **14**: 8-9; **TAP 15**: 9; **16**: 10; **17**: 11; **18**: 12	The ITBS expanded standard score is continuous with Tests of Achievement and Proficiency, 1978, for specific tests. 1982 norms are available for the ITBS Levels 5-14, and TAP. The publisher recommends testing no more than two levels out of level for functional level testing.
Kaufman Test of Educational Achievement, 1985 American Guidance Service	NA	Nov 3 (1-12) Apr 27 (1-12)	**Wide Range Test**: 1-12	Norms were developed using the Rasch model. "Brief" and "comprehensive" forms exist and are equated. Hand-scorable only.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
KeyMath Diagnostic Arithmetic Test, 1971-76 American Guidance Service	N/A	Oct 15 (2-6) Apr 15 (2-6)	**Wide Range Test**: 2-6	Empirical fall and spring norms (developed in 1977-78) for grades 2-6 are available in Supplementary norm tables, which must be requested from the publisher. Directions for interpolating norms are available from the publisher.
Metropolitan Achievement Tests, Diagnostic Battery, 1986 Forms L & M Psychological Corp.	Scaled score	Oct 15 (1-9) Apr 25 (K-9)	**Primer**: K.5-1.4; **Primary 1**: 1.5-2.4; **Primary 2**: 2.5-3.4; **Elementary**: 3.5-4.9; **Intermediate**: 5.0-6.9; **Advanced 1**: 7.0-9.9	Nationally normed, criterion-referenced test battery. Norms and content are coordinated with MAT6 Survey tests.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Metropolitan Achievement Tests, Survey Battery, 1985 Forms L & M **Metropolitan Readiness Tests, 1986** Psychological Corp.	Scaled score	Sept 30 (PreK-1) Oct 15 (K-12) Jan 30 (PreK-1) Apr 25 (K-12) Apr 30 (PreK-1)	**Metro 1**: PreK-K.5; **Metro 2**: K.5-1.5; **Preprimer**: K.0-K.4; **Primer**: K.5-1.4; **Primary 1**: 1.5-2.4; **Primary 2**: 2.5-3.4; **Elementary**: 3.5-4.9; **Intermediate**: 5.0-6.9; **Advance 1**: 7.0-9.9; **Advance 2**: 10.0-12.9	This is the survey battery for the 6th edition of the Metropolitan Achievement Tests. Norms and content are coordinated with the MAT6 Diagnostic Battery.
National Tests of Basic Skills, 1985 Form 1 American Testronics, Inc.	Equal interval score	Oct 17 (K-12) Apr 24 (K-12)	**P**: PreK-K.5; **A**: K.0-K.9; **B**: K.6-1.5; **C**: 1.0-1.9; **D**: 1.6-2.9; **E**: 2.6-3.9; **F**: 3.6-4.9; **G**: 4.6-5.9; **H**: 5.6-6.9; **I**: 6.6-7.9; **J**: 7.6-8.9; **K**: 8.6-10.9; **L**: 10.6-12.9; **M**: 11.6-College	The National and Comprehensive Assessment Program Achievement Series (CAP ACH) have equated scores. The publisher recommends testing no more than one level out of level for functional level testing.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Nelson Reading Skills Test, 1977 Forms 3 & 4 Riverside Publishing Co.	Grade equivalent	Oct 28 (3-9) Mar 8 (3-9)	A: 3-4; B: 5-6; C: 7-9	The publisher recommends testing no more than two levels out of level for functional level testing. Multilevel booklets allow for a wide range of abilities. Group administered individual testing.
Peabody Individual Achievement Test, 1970 American Guidance Service	N/A	Mar 15 (K-12)	A: K-1; B: 1-2; C: 2-3; D: 4-6; E: 7-9	Supplementary norms tables are available from the publisher.
PRI Reading Systems, 1980 CTB/McGraw-Hill	Scale score See comments	See comments	A: K-1; B: 1-2; C: 2-3; D: 4-6; E: 7-9	Scores are linked to the CTBS/U, CAT/C and CAT/E. Test administration dates and norms should correspond to the calendar dates for the appropriate test listed above.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Reading Yardsticks, 1981 Riverside Publishing Co.	Standard score	See comments	**6**: K; **7**: 1; **8**: 2; **9**: 3; **10**: 4; **11**: 5; **12**: 6; **13**: 7; **14**: 8	Norm-referenced score estimates for comparable subtests on the ITBS, Gates-MacGinitie and The 3-R's are available from the publisher for in-level testing only. Norms were developed by equating with the ITBS and TAP.
Scan-Tron Reading Tests, 1985 Scan-Tron Corp.	Not available	Oct 15 (3-8) Apr 23 (3-8)	**8**: 3; **9**: 4; **10**: 5; **11**: 6; **12**: 7-8	Norms were developed by equating with the CAP ACH (1980). In-level and out-of-level norm booklets are available. Tests can be scanned, scored, and results reported through a microcomputer software program.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Sequential Tests of Educational Progress, Series III, 1979 Forms X & Y **CIRCUS, 1972-79** CTB/McGraw-Hill	Standard score	Oct 5 (3-12) May 10 (K-12) Oct 15 (K-3) Jan 15 (PreK)	**CIRCUS A**: PreK-K.5; **B**: K.5-1.5; **C**: 1.5-2.5; **D**: 2.5-3.5; **STEP E**: 3.5-4.5; **F**: 4.5-5.5; **G**: 5.5-6.5; **H**: 6.5-7.5; **I**: 7.5-10.5; **J**: 10.5; 12.9	The CIRCUS is continuous with the STEP. Out-of-level norms are included in the norms booklet. The publisher recommends testing no more than one level out of level.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
SRA Achievement Series, 1978-1985 Forms 1 & 2 Science Research Associates, Inc.	Growth scale value	Oct 4 (1-12) Apr 11 (K-12) Oct 1 (K-12) Apr 22 (K-12) (for 1978 edition)	**A**: K.5-1.5; **B**: 1.5-2.5; **C**: 2.5-3.5; **D**: 3.5-4.5; **E**: 4.5-6.5; **F**: 6.0-8.5; **G**: 8.0-10.5; **H**: 9.0-12.9	For Chapter 1 students, the publisher suggests the following use of test levels: **A**: K, 1; **B**: 2; **C**: 3; **D**: 4; **E**: 5-6; **F**: 7-8; **G**: 9; **H**: 10-12 Quarter-month empirical and interpolated norms are available from the publisher's scoring service. Growth scale value is equivalent to that of the Survey of Basic Skills.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Stanford Achievement Test, 1982, 1986 Forms E & F Psychological Corp.	Scaled score	**1982**: Oct 7 (K-12) Feb 3 (K-1) May 5 (K-12) **1986**: Oct 7 (K-12) May 5 (K-12)	**SESAT 1**: K.0-K.9; **SESAT 2**: K.5-1.9; **Prim.1**: 1.5-2.9; **Prim.2**: 2.5-3.9; **Prim.3**: 3.5-4.9; **Inter.1**: 4.5-5.9; **Inter.2**: 5.5-7.9; **Adv.**: 7.0-9.9; **TASK 1**: 8.0-12.9; **TASK 2**: 9.0-13	Week of testing interpolated norms tables are available from the publisher. 1986 norms (Plus edition) available. Linked to Stanford Diagnostic Tests through Rasch equating. The Stanford is continuous with the Stanford Early School Achievement Test and the Stanford Test of Academic Skills. Writing sample is part of the Stanford.

140

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Stanford Diagnostic Mathematics Test, 1984 Forms G & H Psychological Corp.	Scaled score	Oct 5 (2-12) May 1 (1-12)	**Red**: 1.6-3.6; **Green**: 4.1-5.6; **Brown**: 6.1-7.6; **Blue**: 8.1-12.6	Week of testing norms available through the publisher's scoring service or tables may be ordered from the publisher. The publisher does not recommend out-of-level testing. The SDMT is linked to the Stanford through Rasch equating with no overlap in content.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Stanford Diagnostic Reading Test, 1984 Forms G & H Psychological Corp	Scaled score	Oct 5 (2-12) May 1 (1-12)	**Red:** 1.8-3.8; **Green:** 4.1-5.8; **Brown:** 6.1-7.8; **Blue:** 8.1-12.8	Week of testing interpolated norms available through the publisher's scoring service or tables may be ordered from the publisher. The publisher does not recommend out-of-level testing. Total reading score is not available. The SDRT is linked to the Stanford through Rasch equating with no overlap in content.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Survey of Basic Skills, 1985 Forms P & Q Science Research Associates, Inc.	Growth scale value and scale score	Oct 4 (1-12) Apr 11 (K-12)	**20**: K.5-1.5; **21**: 1.5-2.5; **22**: 2.5-3.5; **23**: 3.5-4.5; **34**: 4.5-6.5; **35**: 6.5-8.5; **36**: 8.5-10.5; **37**: 9.0-12.9	Scores for the reference materials test can be determined with the answer sheet edition only. Quarter month empirical and interpolated/ extrapolated norms are available from publisher's scoring service or may be computed. SBS is equated to the SRA Achievement Series.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
3-R's Test, 1982 Forms A, B & C Riverside Publishing Co.	Expanded standard score	Oct 27 (K-12) Apr 28 (K-12)	**6**: K; **7**: 1; **8**: 2; **9**: 3; **10**: 4; **11**: 5; **12**: 6; **13**: 7; **14**: 8; **15/16**: 9-10; **17/18**: 11-12	The publisher recommends testing no more than two levels out of level for the Achievement Edition. Achievement/ Abilities edition is for levels 9-17/18. The Class Period edition provides national forms for only a single composite score of reading, mathematics and language.
Wide Range Achievement Test, Revised, 1984 Jastak Associates, Inc.	N/A	See comments	**Level I**: ages 5.0-11.11 **Level II**: ages 12.0-adult	Norms are based on age, not grade level and are available in print. New standardization based on Rasch model.

Test Series Year, Form	Expanded Score	Empirical Norm Dates (Appropriate Grades)	Publisher Recommended In Level Grade Ranges (Level: Grade)	Comments
Woodcock Reading Mastery Tests Revised, 1987 Forms G & H American Guidance Service	Not applicable	See comments	**Wide Range Test**: K-12	Continuous norms developed using the Rasch-Wright model are available.
Woodcock Reading Mastery Tests, 1973-78 Forms A & B American Guidance Service	Mastery score (1973) None (77-78)	Oct 15 Apr 15 (1977-78 norms: 2-6) May 15 (1973 norms: K-12)	**Wide Range Test**: K-12	1973 norms are included in the manual. 1977-78 norms tables are available from the publisher.

Names and addresses of major test publishers

154

American Guidance Service
Publisher's Building
Circle Pines, MN 55014-1796
(800) 328-2560
(800) 247-5053

American Testronics, Inc.
P.O. Box 2270
Iowa City, IA 52244
(800) 553-0030
(319) 351-9086

The College Board
45 Columbus Avenue
New York, NY 10023
(212) 713-8000

CTB/McGraw-Hill
Del Monte Research Park
2500 Garden Road
Monterey, CA 93940
(800) 538-9547
(408) 649-8400

Jastak Associates, Inc.
1526 Gilpin Avenue
Wilmington, DE 19806
(800) 221-9728

Psychological Corporation
555 Academic Court
San Antonio, TX 78204-0954
(800) 228-0752
(512) 299-1061

The Riverside Publishing Company
8420 Bryn Mawr Avenue
Chicago, IL 60631
(800) 323-9540
(312) 693-0040

Scan-Tron Corporation
1361 Valencia Avenue
Tustin, CA 92680-6463
(714) 259-8887

Science Research Associates, Inc.
155 North Wacker Drive
Chicago, IL 60606
(800) 621-0476
(312) 984-7000

A glossary of measurement terms

A

achievement test -- an objective examination that measures educationally relevant skills or knowledge about such subjects as reading, spelling, or mathematics.

age norms -- values representing typical or average performance of people of age groups.

average -- a statistic that indicates the central tendency or most typical score of a group of scores. Most often average refers to the sum of a set of scores divided by the number of scores in the set.

B

battery -- a group of carefully selected tests that are administered to a given population, the results of which are of value individually, in combination, and totally.

C

ceiling -- the upper limit of ability that can be measured by a particular test.

criterion-referenced test -- a measurement of achievement of specific criteria or skills in terms of absolute levels of mastery. The focus is on performance of an individual as measured against a standard or criterion rather than against performance of others who take the same test, as with norm-referenced tests.

D

diagnostic test -- an intensive, in-depth evaluation process with a relatively detailed and narrow coverage of a specific area. The purpose of this test is to determine the specific learning needs of individual students and to be able to meet those needs through regular or remedial classroom instruction.

domain-referenced test -- a test in which performance is measured against a well-defined set of tasks or body of knowledge (domain). Domain-referenced tests are a specific set of criterion-referenced tests and have a similar purpose.

G

grade equivalent -- the estimated grade level that corresponds to a given score.

I

informal test -- a nonstandardized test that is designed to give an approximate index of an individual's level of ability or learning style; often teacher-constructed.

inventory -- a catalog or list for assessing the absence or presence of certain attitudes, interests, behaviors, or other items regarded as relevant to a given purpose.

item -- an individual question or exercise in a test or evaluative instrument.

N

norms -- performance standards that are established by a reference group and that describe average or typical performance. Usually norms are determined by testing a representative group and then calculating the group's test performance.

normal curve equivalent -- standard scores with a mean of 50 and a standard deviation of approximately 21.

norm-referenced test -- an objective test that is standardized on a group of individuals whose performance is evaluated in relation to the performance of other individuals; contrasted with criterion-referenced test.

O

objective percent correct -- the percent of the items measuring a single objective that a student answers correctly.

P

percentile -- the percent of people in the norming sample whose scores were below a given score.

percent score -- the percent of items that are answered correctly.

performance test -- designed to evaluate general intelligence or aptitudes. Performance tests usually consist primarily of motor items or perceptual items because verbal abilities play a minimal role.

published test -- a test that is publicly available because it has been copyrighted and published commercially.

R

rating scales -- subjective assessments made on predetermined criteria in the form of a scale. Rating scales include numerical scales or descriptive scales. Forced choice rating scales require that the rater determine whether an individual demonstrates more of one trait than another.

raw score -- the number of items that are answered correctly.

reliability -- the extent of which a test is dependable, stable, and consistent when administered to the same individuals on different occasions. Technically, this is a statistical term that defines the extent to which errors of measurement are absent from a measurement instrument.

S

screening -- a fast, efficient measurement for a large population to identify individuals who may deviate in a specified area, such as the incidence of maladjustment or readiness for academic work.

specimen set -- a sample set of testing materials that are available from a commercial test publisher. This may include a complete individual test without multiple copies or a copy of the basic test and administration procedures.

standardized test -- a form of measurement that has been normed against a specific population. Standardization is obtained by administering the test to a given population and then calculating means, standard deviations, standardized scores, and percentiles. Equivalent scores are then produced for comparisons of an individual score to the norm group's performance.

standard scores -- a score that is expressed as a deviation from a population mean.

stanine -- one of the steps in a nine-point scale of standard scores.

V

validity -- the extent to which a test measures what it was intended to measure. Validity indicates the degree of accuracy of either predictions or inferences based upon a test score.

101

Acknowledgements

This report reflects the thoughts and opinions of many people. The discussions that we had while developing this report helped us to focus on what is important to local school districts and to present concise and accurate information.

Particular thanks are due to the following individuals for their ideas, suggestions, criticisms, and feedback.

Judith Arter, Northwest Regional Education Laboratory
Ronald Boyd, American Institutes for Research
Carolyn Bocella Bagin, American Institutes for Research
Kathy Dorko, American Institutes for Research
Thomas Eissenberg, American Institutes for Research
Chester Finn, Vanderbilt University
Valeria Ford, District of Columbia Public Schools
Steve Frankel, Montgomery County (MD) Public Schools
Pamela Getson, Children's Hospital National Medical Center
Norman Gold, District of Columbia Public Schools
David Goslin, American Institutes for Research
Allan Hartman, Massachusetts State Department of Education

162

Joan Herman, Center for the Study of Evaluation
Elizabeth Heins, Stetson University
M. Kevin Matter, Cherry Creek (CO) Public Schools
Dan Koretz, RAND Corporation
Robert Krug, American Institutes for Research
Bruno Manno, U.S. Department of Education
Doris Redfield, U.S. Department of Education
Janice Redish, American Institutes for Research
Morris Jack Rudner, New Milford (CT) Public Schools
Jeffry Schiller, U.S. Department of Education
Gayle Schindler, American Institutes for Research
Sarah Spatt, American Institutes for Research
Robert Stonehill, U.S. Department of Education
George R. Wheaton, American Institutes for Research
Lauress Wise, American Institutes for Research

183

Index

A

Absences 122
Academic achievement i, 41
Accountability purposes 53
Accuracy 34, 66, 160
Achievement i, ii, iv, ix, x, 1-5, 9-12, 17-19, 23, 27, 28, 32, 34, 36, 39, 41, 48, 53, 54, 57, 61, 62, 72-76, 82, 98, 101, 105, 107, 109, 120, 126, 131, 133, 136, 137, 139, 140, 143, 144, 147, 148, 155, 156
Achievement test i, 10-12, 32, 140, 144, 148, 155
Age norms 155
Appropriate uses 16-18, 43
Average 20, 23, 26, 28-32, 37, 40, 41, 43, 83, 84, 97, 102, 103, 104, 121, 124, 127-129, 155, 157

B

Battery 11, 36, 38, 138, 139, 155
Buros Institute ii, 1, 3, 4, 61, 108, 110, 116

C

Ceiling 156
Compensatory education 5, 77, 81, 82
Counseling 113, 115, 125
Criterion referenced 17, 18, 136
Curriculum ix, 4, 5, 13, 14, 17, 18, 41, 48, 58, 61-63, 67, 70, 71, 75, 91, 93, 99
Custom-made standardized tests ix, 5
Customized test 71, 72, 74
Cut-off score 79

D

Diagnostic test 156
Diversity 123
Domain-referenced test 156

E

Environment 49, 127
Error of measurement 48, 49

F

Finding information x, 105, 107

105

G

General instruction 3, 54
General progress 13
Glossary x, 105, 155
Grade equivalents 33, 69

I

In-level testing 80, 141
In-service programs 104, 124
Informal test 157
Inventory 135, 157
Issues iii, 45, 58, 59, 76, 86, 115, 118-120
Item tryout 12

L

Limitations v, 2, 10, 13, 16, 24, 26, 29, 31, 34, 36, 42
Limited English speaking students 110

M

Match 2, 11, 22, 54, 58, 62-66
Mental Measurement Yearbook 110, 116
Misuse 48, 49
Multiple assessments 79

N

National norms i, 4, 13, 40, 67
National sample 4, 20, 61, 63
Normal curve equivalents 19, 38, 69
Norm referenced 17

O

Objective percent correct 25, 158
Online information retrieval systems 107, 111
Out-of-level testing 80, 145, 146

P

Participation 122, 123
Percent correct 24, 25, 29, 50, 158
Percentiles 12, 19, 27-29, 39, 40, 135, 160
Performance test 158
Press x, 2, 6, 47, 85-87, 89-94, 108, 109
Programs for students 125
Publishers x, 4, 12-15, 35, 40, 66, 80, 105, 107, 108, 111, 120, 151

R

Rating scales 159
Raw scores 19-21, 23, 24, 28, 30, 33, 35, 37, 41, 43, 50
Reliability 73, 159

10

Reliable 78, 81, 98
Reporting i, ii, 1, 36, 41, 72, 90
Results i, ii, iii, iv, v, 1, 2, 11, 17, 22, 31, 35, 37, 41, 69, 73, 74, 75, 79, 84, 90-93, 95-102, 116, 126, 141, 155

S

Screening 133, 159
Selecting students 79, 83
Single test score 49
Specific skills 13, 14
Specimen set 159
Stability 124
Staff development 124
Staff experience 124
Standard scores 19, 34-36, 38, 136, 157, 160
Standardized tests i, iii, iv, v, ix, x, 1-5, 12, 45, 47, 49, 53, 54, 56, 57-59, 61, 62, 69, 76, 77, 79, 80, 90, 105, 107, 122, 126, 158, 160
Stanines 19, 30, 31, 39, 69
Student
  gains 41
  performance i, iii, 22
  population 11, 123
Study skills 99, 125

T

Table of specifications 64
Test
  critiques 111

development 112
preparation 4, 57, 74, 75
results i, ii, iv, v, 1, 2, 31, 41, 73-75, 90-93, 95, 99, 116
reviews 107, 108, 110, 111
score iv, 21, 23, 25, 32, 48, 49, 51, 55, 74, 78, 83, 160
selection i
specifications 10
title 109, 111
Test-taking skills 54, 57, 58
Tests in Print 108
Total percent correct 24, 29, 50

V

Valid indicators 78
Validity 5, 72, 73, 160
Variability 12, 15, 23